

A Thesis Submitted for the Degree of PhD at the University of Warwick

Permanent WRAP URL:

<http://wrap.warwick.ac.uk/137779>

Copyright and reuse:

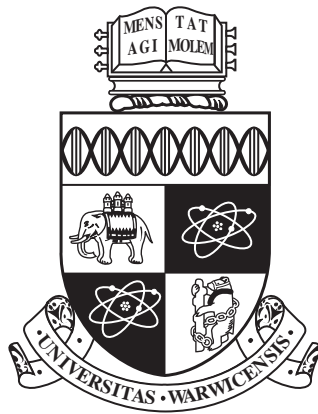
This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it.

Our policy information is available from the repository home page.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk



**Studies in Urban Informatics: Tools and
Techniques to Explore Socio-Ecological Urban
Systems**

by

Nicholas E. Johnson

A thesis submitted to The University of Warwick

in partial fulfilment of the requirements

for admission to the degree of

Doctor of Philosophy

Department of Computer Science

The University of Warwick

December 2018

Abstract

This work details the emerging discipline of urban informatics and the diverse set of tools, data and techniques necessary for the quantitative analysis of urban systems. Three studies are presented based on distinct data types including administrative data, user-generated data and sensor data. The first study focuses on urban waste management and demonstrates how existing administrative datasets can be used to forecast waste generation, which can be useful for optimizing waste collection efforts and developing future waste reduction strategies. The second study shifts from administrative data to focus on the benefits of public participation and the challenges of working with user-generated data. The final study presents a multi-week calibration campaign to evaluate calibration techniques and the quality of data generated by a low-cost air quality monitoring platform in order to increase the spatial resolution of PM_{2.5} measurements in an urban environment. The results from and evaluation of these studies highlight the potential for these urban data streams to provide new insight into socio-ecological urban systems, and create new opportunities for local governments to operate in a more effective, efficient and sustainable way.

This work is dedicated to my wife Cristina, daughter Anna and son Leonardo
for their endless love, support and encouragement.

Acknowledgements

This work has been the result of my experience living in New York City and the numerous relationships that have supported, inspired and encouraged my efforts. First and foremost, I thank my advisors Stephen Jarvis and Constantine Kontakosta. Each was incredibly generous and supportive throughout my studies and I greatly appreciate their comments, advice and direction, which influenced this thesis.

I am very grateful to Francois Grey who has been an incredible mentor, and taught me the value of collaboration and the importance of being open-minded. Indeed, this work would not have been possible without his guidance and encouragement, and I cannot thank him enough. I would like to thank Tom Igoe and Dan O’Sullivan who revolutionized my thinking about society and technology, and who were formative in my passion for trash. I also thank Manuela Veloso for her incendiary enthusiasm for science and discovery; it has been a privilege to know her.

At the Center for Urban Science and Progress, where I conducted my research, I was fortunate to know Steve Koonin, Michael Holland, Pat Bowers, Jennifer Cho, Christine Battaglia, Andy Karpf, Federica Bianca, Theo Damoulas, Ravi Shroff, Justin Salamon, Boyeong Hong, Bartosz Bonczak and the countless others who not only shared their knowledge and experiences, but embraced my obsession with waste. In particular, I would like to thank Greg Dobler who was incredibly generous with his intellect and moral support, and from whom I learned scientific rigor, as well as Charlie Mydlarz and Mohit Sharma who helped me overcome numerous technical obstacles and who challenged me with their own ideas as we developed solutions together.

I wouldn’t be here without my loving family: my father David, my mother Julia, my sisters Carrie and Katherine and their families, who have always been

a source of support and inspiration; and my extended Focolare family who have shaped my view of the world, especially the Focolarini in New York City who give selflessly to no end. But most of all, I thank my wife Cristina, who is the love of my life and an amazing mother to our children. Her compassion and love for those around her continues to inspire me.

Declarations

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree. The work presented (including data generated and data analysis) was carried out by the author and parts of this thesis have been published by the author:

- Johnson, Nicholas E., Bonczak, B. and Kontokosta C. E. “Using a gradient boosting model to improve the performance of low-cost aerosol monitors in a dense, heterogeneous urban environment.” *Atmospheric Environment* 184 (2018): 9-16.
- Johnson, Nicholas E., et al. “Patterns of waste generation: A gradient boosting model for short-term waste prediction in New York City.” *Waste Management* 62 (2017): 3-11.
- Johnson, Nicholas E., and Francois Grey. “Landfill Hunter: Learning about Waste through Public Participation.” *Human Computation* 3.1 (2016): 243-252.

The author also contributed to other published works, which were not directly included in this thesis:

- Traunmueller, M., Johnson, N., Malik, A., and Kontokosta, C. E. “Digital footprints: Using WiFi probe and locational data to analyze human mobility trajectories in cities.” *Computers, Environment and Urban Systems* (2018)
- Kontokosta, Constantine E., Hong, B., Johnson, N. E., and Starobin, D. “Using machine learning and small area estimation to predict building-level municipal solid waste generation in cities.” *Computers, Environment and Urban Systems* 70 (2018): 151-162.

-
- Kontokosta, C. E., and Johnson, N. “Urban phenology: Toward a real-time census of the city using Wi-Fi data.” *Computers, Environment and Urban Systems*, 64 (2017): 144-153.
 - Traunmueller, M., Johnson, N., Malik, A., and Kontokosta, C. E. “Digital Traces: Modeling Urban Mobility using Wifi Probe Data.” 6th International Workshop on Urban Computing ACM KDD, Halifax, Nova Scotia, Canada (2017)
 - Kontokosta, C. E., Johnson, N., and Schloss, A. “The Quantified Community at Red Hook: Urban Sensing and Citizen Science in Low-Income Neighborhoods.” (2016) arXiv preprint arXiv:1609.08780.

Sponsorship and Grants

The research presented in this thesis was made possible by the support of the following benefactors and sources:

- UK Engineering and Physical Sciences Research Council (EPSRC) Centre for Doctoral Training in Urban Science and Progress (EP/L016400/1)
- New York University's Center for Urban Science and Progress

Abbreviations

ACS Amerian Community Survey

DEC Department of Environmental Conservation

DIY do-it-yourself

DSNY New York City Department of Sanitation

ECHO Enforcement Compliance History Online

EPA Environmental Protection Agency

GBRT Gradient Boosting Regression Tree

ICTs information communication technologies

LEHD Longitudinal Employer-Household Dynamics

LOD limit of detection

LPO Low Pulse Occupancy

PLUTO Primary Land Use and Tax Lot Output

MAE mean absolute error

MGP metal, glass and plastic

ML machine learning

MSW municipal solid waste

NYC New York City

OLS Ordinary Least Squares

PPD42 Shinyei PPD42

R² the coefficient of determination

RCRA Resource Conservation and Recovery Act

RMSE root mean squared error

RMSLE root mean squared logarithmic error

RSS residual sum of squares

TEOM tapered element oscillating microbalance

US United States

Contents

Abstract	ii
Dedication	iii
Acknowledgements	iv
Declarations	vi
Sponsorship and Grants	viii
Abbreviations	ix
List of Figures	xvi
List of Tables	xvii
1 Introduction	1
1.1 Thesis Contributions	6
1.2 Thesis Structure	10
2 Background Research	12
2.1 Urban Data Taxonomy	12
2.2 Administrative Data for Modeling Waste Generation	16
2.3 User-Generated Data for Education and Landfill Monitoring	18
2.4 Sensor Data for Monitoring Air Quality	20
2.5 Chapter Summary	22
3 Investigating Urban Data: A Gradient Boosting Model for Short-term Waste Prediction in New York City	23
3.1 Waste in NYC	24
3.2 DSNY data	25

3.3	Data Collection	26
3.4	Methods	28
3.4.1	Model Building	28
3.4.2	Feature Engineering	30
3.5	Results and Discussion	33
3.5.1	Model Performance	33
3.5.2	Feature Importance	36
3.5.3	Detailed Discussion of Findings	37
3.6	Discussion of Administrative Data	40
3.7	Chapter Summary	42
4	New Urban Data Streams: The Role of Public Participation in Urban Data	45
4.1	Landfills in the United States	46
4.2	Methodology	48
4.2.1	Application Design	48
4.2.2	Data Assessment and Validation	51
4.3	Results and Discussion	52
4.3.1	Data Contributions	52
4.3.2	Offline Outcomes	55
4.3.3	Challenges and Limitations	57
4.4	Discussion of User-generated Data	59
4.5	Chapter Summary	60
5	Evolving Technologies for Data Acquisition: The Development and Evaluation of Commodity Hardware for Monitoring PM2.5	63
5.1	Methods	64
5.1.1	Node Design	64
5.1.2	Reference Instrument	65
5.1.3	Study Location	66
5.1.4	Performance Evaluation	66

5.1.5	Calibration Approaches	68
5.2	Results and Discussion	70
5.2.1	Ambient Conditions	72
5.2.2	Limit of Detection	75
5.2.3	Calibration Results	75
5.2.4	Main findings	77
5.2.5	Limitations	79
5.3	Discussion of Sensor Data	81
5.4	Chapter Summary	83
6	Discussion: The Dynamics of Data in Urban Informatics Research	86
6.1	Towards an Ecological Urban Informatics	87
6.2	Partnerships and Collaboration	91
6.3	Complementary Roles of Urban Data	92
6.4	Data Quality	94
6.5	Data Privacy	95
6.6	Limitations and Challenges of the Case Studies	98
6.7	Recommendations for Local Governments and Practitioners . . .	100
6.8	General Discussion	101
7	Conclusion	104
7.1	Future Work	105

List of Figures

3.1	New York City’s weekly refuse and recycling collection tonnages from 2004-2015. The vertical red lines indicate significant winter weather events with above average snowfall.	26
3.2	Daily per capita distribution of waste and recycling for New York City Department of Sanitation (DSNY) sections. The black lines indicate a normal distribution.	27
3.3	Autocorrelation with 95 confidence intervals shown in light blue where the standard deviation is computed using Barlett’s formula.	31
3.4	DSNY waste collection map. Large multi-colored areas indicate a DSNY section while smaller multi-colored areas reflect census blocks.	34
3.5	Spatiotemporal refuse prediction integrated for all sections using multiple feature groups.	35
3.6	Spatiotemporal refuse prediction for an individual section in Manhattan.	36
3.7	Importance ranking for internal and external features used in refuse prediction.	37
4.1	Long-haul truck dumping waste into a landfill. Screenshot taken from Google Maps.	49
4.2	Landfill images used as examples for users prior to starting. Images were taken from Google Maps.	50
4.3	Descriptive analysis of user participation rates. Figure 4.3A describes the distribution of task responses per user and Figure 4.3B shows the distribution of time spent per task.	54

4.4	Example user responses aggregated by task. Figure 4.4A highlights user ability to identify similar landfill definitions while figure 4.4B shows the results plotted with Google Maps.	55
5.1	Meteorological measurements taken from La Guardia airport over the study period. (a) Temperature (blue) and sea level pressure (red), (b) precipitation (blue) and humidity (red line), and (c) wind speed (blue line) and wind direction (red points).	67
5.2	Pairwise plots between three Shinyei PPD42 (PPD42) devices and a reference tapered element oscillating microbalance (TEOM) based on hourly data collected from February 7th 2017 to March 25th 2017.	71
5.3	Linear model fit for hourly data collected from three PPD42 sensors and a TEOM reference monitor between February 7th 2017 and March 25th, 2017.	73
5.4	Feature importance for the ridge regression model (A) and the gradient boosting regression model (B).	76
5.5	Scatter plots of three Shinyei PPD42 sensors calibrated with three different techniques. Sensors are calibrated through a multi-linear regression, ridge regression and gradient boosting regression tree model.	77
5.6	Comparison of calibration results with a reference instrument using different calibration techniques including multiple linear regression, ridge regression and gradient boosting regression tree models. Hourly PM2.5 measurements were obtained from three Shinyei PPD42 sensors co-located with a TEOM reference instrument from February 7th through March 25th, 2017.	78

6.1	A map of New York City’s waste disposal network. The orange lines represent possible routes from DSNY transfer stations to landfills located in states known to accept NYC waste. Specific routes were generated by the Google Map API in order to provide specific truck driving routes.	89
-----	---	----

List of Tables

2.1	A taxonomy of urban data according to how the data was generated.	14
3.1	Complete list of features used for prediction	32
3.2	Model prediction results	35
4.1	A description of the data generated by the Landfill Hunter application.	53
5.1	Results of a sensitivity test to evaluate the relationship between meteorological conditions and the Shinyei PPD42 sensor response.	74
5.2	Results from calculating the lower limit of detection for the PPD42 during a field calibration campaign with a TEOM reference instrument. Units are in $\mu\text{g}/\text{m}^3$	75
5.3	Comparison of results from three calibration techniques.	76

CHAPTER 1

Introduction

Today, over half of the world's population lives in cities and by 2050 nearly two thirds of the global population will live in urban areas [89]. As a result of this rapid urbanization, cities around the world are fundamentally changing the natural landscape, which has important consequences for both local and global ecosystems [39]. At the local level, cities generate an enormous volume of pollution that can lead to problems such as poor air quality, contamination of fresh water and reduced biodiversity. At the global scale, urban centers are primary sources of greenhouse-gas emissions making them one of the main drivers of global climate change [39]. It is now well understood that the future of Earth's ecosystem is increasingly influenced by human activities at the urban scale and it is therefore essential to understand the complex and dynamic relationship between cities and nature in order to ensure a sustainable future [2, 3, 80].

While our current understanding of urban dynamics has evolved dramatically, existing theories and methods for studying cities are rooted in the growth and development of urban areas during the 19th century. Economic growth, technological advances and new forms of transportation brought about by the industrial revolution resulted in the dramatic growth and expansion of cities. This rapid urbanization, combined with changes in manufacturing and industrial processes, led to the deterioration of the urban environment, especially for working residents who were exposed to horrendous living conditions including overcrowding, excessive pollution and substandard housing. In response to these challenges, cities began to develop new strategies for planning and designing urban areas, which ultimately led to the development of the discipline of urban planning. Broadly, urban planners sought to create master plans to spatially

organize cities in order to create spaces that combined the benefits of both town and country living. These top-down design approaches were largely based on theories and models developed at that time to describe and explain the underlying principles that govern cities. The application of these urban theories and the development of urban modeling throughout the 19th and 20th centuries not only provided the groundwork for understanding cities and urban activity, but laid the foundation for the cities of today.

Broadly, an urban theory is an explanation of an urban aspect or phenomena that can be sufficiently validated through the process of making predictions. In order to evaluate and assess the validity of a theory, the theory is often translated into a mathematical or logical model, which can then be tested and simulated repeatedly. Urban modeling therefore, is the process of defining, building and applying various urban theories in order to test and assess the validity of a theory, often for the purposes of forecasting, designing and planning [9].

The primary urban theories that formed our current understanding and approach towards the scientific study of cities are based broadly on three concepts developed during the late 19th century and early 20th century. These theories include economic location theory, social physics, and spatial morphology [3]. Economic location theories generally propose solutions that, in practice, enable businesses, industries and households to locate in space based on a predetermined set of constraints [9]. These constraints primarily focus on economic decision-making related to the spatial distances between locations, but also include examples of macro- and micro-economic theories that describe spatial patterns resulting from land use value, relationships between modes of production and consumption, and the effects of competition for land use. Social physics theories tend to focus on understanding human behavior through the application of theories from physics, and the discovery of patterns in human activity that can be described mathematically. These theories argue that the social distribution of human behavior such as social interactions and population movement, can be described by mathematical formulations and various gravitational and

energy models. Theories related to spatial morphology are based on identifying patterns that describe how the physical infrastructure and land use of cities are geographically and spatially structured. Emphasis is placed on describing the relationship between urban elements (e.g. covered and open land), identifying elements that provide linkages to various parts of the city, and examining the spatial distribution of urban services for human usage.

Early in the 20th century, these concepts, theories and models became the principal tools and techniques actively employed by planners to shape cities. Using these tools, urban planners focused on zoning and land use to centralize and control industrial and manufacturing processes in order to control pollution generation and preserve green and open spaces, which ultimately led to the expansion of the urban periphery, and the creation of decentralized and low-density suburb communities. Despite the initial momentum for and advances in city planning, however, urban theories and modeling concepts about cities began to change dramatically toward the end of the 20th century. Contrary to traditional urban planning theories that claimed cities could be explained by simple patterns of movement and organization, cities increasingly appeared to be unorganized, unruly and unpredictable. Transportation and public housing, in particular, continued to worsen and traditional strategies of urban planning seemed to become increasingly ineffective in addressing these issues. Indeed, top-down planning approaches often seemed to make urban problems worse rather than leading to improvements, and ultimately led to the notion that urban planning itself, as well as the urban theories that motivated new planning approaches, were the underlying problem.

Critiques of these urban planning approaches focused on flaws in the underlying urban theories. First, urban theories were largely based on the assumption that cities maintain a state of equilibrium, which not only failed to acknowledge that cities are influenced by numerous internal and external factors, and do not exist in isolation, but also that cities do not automatically return to a state of equilibrium when disrupted [10]. Similarly, cities were often considered to be

a collection of structured units with explicit functions that could be organized and controlled from the top down. Planners also failed to consider cities as being potentially unpredictable and capable of generating surprising outcomes.

The unintended consequences generated by the application of flawed urban theories and urban modeling attempts, led to important shifts in how cities were conceptualized and how urban modeling was approached. Fundamentally, the focus shifted away from viewing the city as a static structure in space, and began to focus on the underlying processes that motivate urban change and evolve over time. Emphasis was placed on understanding the growth and behavior of cities, similar to that of an organism, and developing more dynamic, disaggregated and bottom up modeling approaches. Simultaneously, these conceptual shifts were accompanied by mathematical advances and increased computation power of computers, which created new possibilities to develop disaggregated models focusing on the interaction between individual system components.

At the beginning of the 21st century, a new framework has emerged motivated by the adoption of new information communication technologies (ICTs) and access to new big data sources of information about urban activity. This framework, known as “Smart Cities”, broadly applies to urban areas with characteristics that include increased use and expansion of networked infrastructure, an emphasis on business-led urban development, an aim for social inclusion, encouragement and proliferation of high-tech and creative industries and an emphasis on a social and environmental sustainable city [16]. Fundamentally, the “Smart City” concept centers on utilizing ICTs and new sources of big data to improve urban operations and deliver efficient and sustainable services. For local and city governments in particular, the “Smart City” framework also includes new opportunities to enhance transparency and accountability.

Numerous cities around the world have adopted this framework through a variety of technological innovations and initiatives. Songdo, South Korea, for example, was one of the first cities to be designed with extensive use of digital infrastructure. Built on reclaimed land, the city became known as a “ubiqui-

tous city” for its extensive networked infrastructure that included sensors to monitor transportation and energy systems, video conferencing capabilities in individual households and high-speed Wi-Fi access throughout the public transportation system. Singapore provides a complementary example because of its increased use of data to address difficult urban challenges. In particular, the city of Singapore has focused on collecting and using mobile phone data to assess and improve public transportation systems. Furthermore, the city continues to promote and invest in an open data platform to enable new analyses and promote innovative solutions to other urban challenges. Rio de Janeiro offers another unique example of a city adopting ICTs to improve urban operations. Specifically, the city has launched a citywide operations center connecting various city agencies allowing for the centralization of real-time information, which enables officials to collaborate and manage city services, as well as coordinate emergency response teams in the event of a crisis or disaster. While the number of cities who have adopted the smart city paradigm in various forms continues to grow, more notable examples include Barcelona, London, Dublin, New York City, Dubai and Tokyo.

Given this context, the discipline of urban informatics has emerged to leverage smart city technologies and new sources of “big data” being generated by cities in order to study complex urban systems [85]. Urban informatics is the study of urban patterns and processes through the integration, analysis, visualization and interpreting of structured and unstructured data with the goals of improving urban operations and resource management, advancing theoretical insight into urban systems including infrastructural, physical and socioeconomic systems, developing programs for improved public engagement and participation, and implementing strategies for long-term planning and impact assessment of urban policies [85]. By integrating a variety of novel data sources including administrative datasets, in situ sensor data and user-generated data, urban informatics research often takes a new data-driven modeling approach for knowledge discovery in which an explanatory or predictive model is derived from the

data. This empirical modeling approach allows for a “bottom-up” understanding of urban phenomena, which differs significantly from the top down modeling approaches traditionally used by urban planners throughout the 20th century. Instead of defining an overarching theory to describe a phenomena and using data to test the theory, this new approach uses the input data to identify patterns and define an appropriate model, which is then used to formulate a new theory. Ultimately, the goal of urban informatics is to generate new insights and actionable intelligence from a variety of new sources of information and analytical methods in order to improve quality-of-life measures that range from public health and safety to sustainability and resilience.

At the turn of the 21st century, cities have already witnessed unprecedented population growth. The monumental challenges created by this rapid population growth have fueled modern efforts to understand cities and the underlying dynamics that govern them. Similar to the industrial cities at the turn of the 19th century, there is dire need to better understand the dynamics of urban systems in order to provide healthy, livable and sustainable cities. Unlike the industrial cities at the turn of the 19th century, however, we now have new digital technologies and the nascent field of urban informatics to enable new approaches towards studying and understanding the complex dynamics that shape our cities.

1.1 Thesis Contributions

Given this context, this thesis evaluates the usefulness of new data sources and tools for data acquisition through a series of empirical urban informatics studies aimed to understand and explore socio-ecological urban systems. Specifically, these studies focus on waste management and air quality, which are not only complex socio-environmental processes well-suited for urban informatics studies, but also fundamental urban challenges with direct impact on residents. These studies were selected to demonstrate diverse examples of urban informatics re-

search based on specific types of urban data and tools for data acquisition in order to provide a definition and overview of urban informatics research that showcases the breadth and depth of the field, as well as describe, compare and contrast new sources of urban data, highlight state-of-the-art methodologies for data collection, analysis and modeling, and identify recurring challenges and opportunities for future work. The studies presented in this thesis are intended to be representative and exemplary urban informatics research.

The first two studies focus on urban waste management and use administrative and user-generated data streams to demonstrate their ability to generate new insights, improve citywide operations and facilitate public participation. Urban waste represents a fundamental and universal urban challenge that not only has a direct impact on local communities and city agencies, but also has important environmental consequences at local, regional and global scales [38]. The third study focuses on urban air quality and the usefulness of low-cost in situ sensors to provide meaningful air quality measurements. Urban air quality is a similar universal urban challenge with significant impacts on local residents, as well as regional and global ecosystems. Though urban waste and urban air quality are two distinct urban processes, they both reveal a relationship between nature and society that is mediated through cities. This socio-ecological relationship is precisely the complex and dynamic phenomena that can be studied and understood in greater detail through the discipline of urban informatics. Indeed, the three studies presented in this thesis integrate novel data sources in order to demonstrate new approaches and data-driven methods to study the interactions, processes and mechanisms that shape and define these socio-ecological urban processes.

The contributions in this thesis include:

- a new method for predicting waste generation using a machine learning model that uses historical municipal solid waste (MSW) collection data supplied by the New York City Department of Sanitation (DSNY) in conjunction with other datasets related to New York City (NYC) to forecast

MSW generation. Traditional methods for forecasting waste generation have largely focused on identifying the underlying factors that influence waste generation and forecasting through regression analysis, time series modeling and system dynamics models. This thesis presents a more advanced machine learning model for predicting waste that incorporates a variety of diverse independent features and provides highly accurate predictions across short time spans. The model could also be adapted for use in similar cities.

- a new understanding of urban waste generation, which was gained through analysis of highly granular historical data and the development and application of a forecasting model. Specifically, this research reveals a regularity of urban waste generation not previously identified, especially at the resolution presented in this research. The identification of this regularity is not only important for predictions across short time periods, but also suggests that even the most rudimentary data about historical waste generation can provide the necessary information for forecasting future waste generation in cities around the world. Furthermore, this research was able to identify new features useful for accurately predicting waste generation including temperature, humidity, precipitation and weather events.
- new insight into the potential learning opportunities and secondary outcomes generated by online crowdsourcing initiatives. Traditionally, crowdsourcing initiatives, both online and offline, focus on developing tools and approaches that optimize user participation (i.e. expedient task completion) and maximize user incentive in order to efficiently analyze or collect the necessary information. While the crowdsourcing project presented in this thesis also sought to achieve these goals, the outcomes revealed new insight for these initiatives to bridge online and offline activity, as well as provide learning opportunities with significant potential to raise awareness and engage citizens in understanding broader urban issues.

- a new source of information about landfills in the United States. For a variety of reasons, there exists a significant lack of information about landfills in the United States and no comprehensive source of data about the specific location, size and composition of these landfills. Through a novel online crowdsourcing initiative, this research is able to solicit public participation to generate and analyze data about landfills to overcome existing knowledge gaps. The resulting dataset provides new information about the specific locations of landfills, the total land area for an individual landfill and the geographic boundaries of these landfills, which can serve a variety of uses such as long-term landfill monitoring, standardized reporting and improved citizen awareness. Importantly, user-generated data identified 729 landfills with a cumulative land area of 576km².
- a new method for integrating commodity hardware components into an in situ platform capable of providing real-time information about PM2.5 levels. Specifically, this research describes the design of a novel hardware platform that not only provides sophisticated data collection, storage and transmission capabilities, but also enables rapid, economic and low-maintenance sensor deployments suitable for large-scale network scenarios.
- a new understanding of the performance of low-cost air quality monitors. A growing body of research is focused on evaluating the performance of commodity hardware through laboratory testing and field calibration campaigns. The work presented in this thesis contributes to that research by providing a new context and environment for sensor evaluation. Specifically, this research evaluates multiple low-cost sensor devices in a dense urban setting under extreme weather conditions for an extended period of time. This new context for evaluation contributed to a new understanding of the performance of these devices in diverse conditions.
- a new method for calibrating a low-cost device by incorporating publicly

available data sources into a robust machine learning model. Traditional calibration approaches have been focused on establishing a linear relationship between a test device and reference instrument, which allows for convenient and simple adjustments to be made on the test device. This research builds upon traditional regression approaches by implementing a machine learning model to calibrate individual devices, and evaluating the results and outcomes of implementing such a method. While this research does demonstrate improved calibration results when implementing a more advanced machine learning model, it also articulates the possible dangers related to using these approaches for calibration.

1.2 Thesis Structure

The structure of this thesis is organized as follows. The next chapter provides specific background information related to the individual studies presented in this thesis and includes an overview of previous work related to predictive models for waste generation, the role of user-generated data related to landfills and work related to air quality monitoring with low-cost sensors. Additionally, this section includes a taxonomy of urban data and frames the selection of the individual studies presented in this thesis.

Chapters 3-5 present individual studies based on three different urban data types. Broadly, each of these chapters begins with initial contextual information including a description of the data being used and the overall objective of the study, followed by a detailed description of the methods and a discussion of the results and findings. While much of this discussion is directly related to the individual study, a general discussion is also included in order to link the individual studies to the broader findings of this thesis.

Chapter 6 includes a detailed discussion of findings that relate more broadly to the field of urban informatics and the thesis as whole. This discussion is largely derived from a collective evaluation of the three individual studies pre-

sented in Chapters 3-5. This chapter discusses topics related to the practical application of urban informatics and the complementary roles of various urban datasets, as well as issues related to data quality and data privacy. The chapter also discusses the challenges and limitations related to the individual studies, but places this discussion in the broader urban informatics context.

Chapter 7 concludes with a summary of the work presented in this thesis and final comments related to the field of urban informatics, as well as a discussion of potential future work.

CHAPTER 2

Background Research

One of the key factors motivating urban informatics research is the widespread adoption of new ICTs, including mobile phone and web-based technologies, which have enabled new sources of “big data” that have the potential to augment traditional datasets and provide new insights into urban systems with increased spatial and temporal resolution. The term “big data” has been used to describe the massive structured and unstructured data created from a variety of sources, such as government surveys, city operations, and social activity, whose complexity and size require new techniques and technologies to extract meaning [85]. These new streams of “big data”, in an urban context, can provide detailed information about urban populations, the built infrastructure and the environment, which are essential for maintaining efficient day-to-day services and identifying and understanding complex urban phenomena.

2.1 Urban Data Taxonomy

Given the diverse sources of information being generated by cities, researchers have begun to develop taxonomies for urban data in order to establish a common framework for research. A variety of urban data taxonomies have been presented but the most common approach to classify urban data is based on how data are being generated. For example, [52] presents three categories of urban data including directed, automated and volunteered data. Directed data are data generated by traditional forms of surveillance or human interaction including CCTV cameras, immigration checkpoints and fingerprints etc. In contrast, automated sources of data refer to transactions and activity using technological systems that will automatically generate information when used, which includes

checkout scanners, mobile devices generating location information and various forms of online activity. Finally, [52] defines volunteered information as information being generated by individuals such as information published to social media platforms or crowdsourced information contributed to OpenStreetMap. An alternative urban data taxonomy has been presented by [60], which classifies data in a similar manner but provides varied terminology and a fourth category. These categories of urban data include transactional data, in situ sensor data, remote sensor data and citizen science data. A taxonomy presented by [85] provides an even more diverse set of categories to distinguish and sub-divide between the broader categories defined by [52] and [60]. [85] group urban data into six categories of sensor systems, user-generated content, administrative data, private sector data, arts and humanities data and hybrid data.

As demonstrated by these existing taxonomies, urban data can be described in a multitude of ways and often consist of overlapping characteristics making a comprehensive taxonomy difficult. However, for the purposes of this thesis, urban data are divided into three categories based on how the source data are generated including administrative data, user-generated data and sensor data. Broadly, these three categories capture the fundamental types of urban data as defined by previous taxonomies, while also sufficiently differentiating between important data characteristics. Using this baseline urban data taxonomy not only allows for the study, comparison and analysis of urban data more generally, but also enables in-depth inspection and evaluation of specific datasets in order to better understand, define and assess the field of urban informatics more broadly. Table 2.1 provides an overview of the different types of urban data based on this taxonomy, and example sources that may generate such data.

Administrative data, the principal source of information about cities, are the text and numerical records generated by government agencies and commercial businesses during their routine course of business [60]. These data often provide basic information about urban infrastructure (e.g. building characteristics, transportation systems, street trees), administrative information (e.g. school

Table 2.1: A taxonomy of urban data according to how the data was generated.

Urban Data Type	Example Sources
Administrative	Transactional (tax records, land use data, waste tonnages), surveys (census records, polls), historical records (annual reports, building permits, budgets)
User-generated	Web-based (web searches, browser history), social media (Twitter, Facebook), citizen science (online crowdsourcing, participatory sensing)
Sensor-based	Remote sensing (satellite imagery, LIDAR), in situ sensing (air/water quality monitoring, video cameras), connected infrastructure (building energy, call-detail records, public Wi-Fi networks)

districts, zoning boundaries), citywide operations (e.g. trash collection, call complaints, budget reports) and population information (e.g. socioeconomic and demographic information). Administrative datasets have been a fundamental element of quantitative urban research since the 1960s and continue to grow as more and more city agencies are being mandated to collect, maintain and disseminate these data in order to improve operational efficiencies and create greater government transparency [7, 61]. In New York City, for example, former Mayor Michael Bloomberg launched the NYC Open Data portal in 2013, which provides unprecedented access to current and historical city records [15].

Despite the increased availability of these administrative datasets, these data are often unable to provide the information necessary to study the complexities of urban systems. These limitations can occur for a variety of reasons including data reliability, bias, completeness and resolution. Furthermore, administrative datasets often only provide static information due to their infrequent collection (>1yr) and release to the public. Efforts to overcome these limitations often focus on developing advanced statistical methods and integrating diverse datasets from a variety of sources.

User-generated data refers to data created by individuals through both active and passive usage of mobile and internet connected technologies [85]. This data not only includes online social media activity (e.g. Facebook, Twitter), web

usage, mobile phone usage and other data generated by daily social activity, but also includes a range of citizen-based participatory sources such as online crowdsourcing. These data are often generated in real-time and include geo-located information, which can be used to understand urban mobility, neighborhood opinions and sentiment and identify patterns of human activity. Additionally, social media data are increasingly being used to create dynamic profiles of individuals and populations in order to provide targeted services and enhanced messaging campaigns. This novel source of information is not only dramatically changing traditional approaches to urban research by augmenting existing administrative datasets and providing new insights into human behavior, but also fundamentally changing the ability of the public to participate and engage in the planning and management of cities by providing new forms of communication and collaboration between public officials and urban residents.

Sensor systems and satellite remote sensing technologies are also foundational sources of information that have provided important data for understanding urban activity and urban ecological systems. Historically, these sources have been based on sparse monitoring networks and remote sensing technologies, which often fail to capture adequate spatial and temporal resolution. Recently, however, in situ sensing has received increased attention because of the increased availability of low-cost sensing technologies and increased Wi-Fi and cellular connectivity throughout major urban centers. This connected infrastructure not only allows for the deployment of dense sensor networks that can generate data with greater spatial and temporal resolution, but also provides new opportunities to integrate sensor technologies into a wide variety of urban infrastructure such as transportation, energy and waste systems [85]. The combination of low-cost sensors and enhanced networking possibilities, creates the potential to augment existing data sources with high resolution information enabling researchers and city officials to monitor and evaluate urban systems in real-time [21].

Collectively, these new data streams can provide unprecedented sources of

information about complex and dynamic urban systems. Combined with new data-driven modeling approaches for knowledge discovery, the discipline of urban informatics offers new possibilities to improve and enhance our understanding of cities, which can lead to improved urban planning and management, create more transparent and efficient city governments and improve quality of life for urban dwellers. Ultimately, cities are experiencing a tsunami of novel data, new technologies and new analytical methods that can combat the current challenges presented by rapid urbanization and lay the foundation for resilient, equitable and sustainable cities in the future.

The rest of this chapter provides background research and previous work for the individual studies presented in this thesis. Specifically, these are divided into techniques used to model and predict waste generation based on administrative data, the use of public participation to validate existing datasets and generate new sources of data, and the potential usefulness and capabilities of using low-cost technologies to provide real-time high resolution air quality measurements.

2.2 Administrative Data for Modeling Waste Generation

With the rapid development of urban environments around the world, municipal waste generation is fast becoming one of the most pressing issues facing cities globally. Currently, at 3.3 million tons per day, the global production of waste is already becoming unmanageable, and this rate is expected to grow to 11 million tons per day by 2100 [42]. Given these trends, effective urban waste management systems are essential, and in order to provide these services in an environmentally sound and financially sustainable way, there is an urgent need for basic understanding of the amount and composition of the materials produced [11, 77]. Furthermore, forecasting waste generation becomes a critical aspect of urban waste management that provides city agencies the ability to optimize collection and disposal operations in the short term, as well as develop

long-term strategies for disposal planning, policy development, and implementation of waste reduction programs [19].

The collection, management, processing and disposal of MSW is an increasingly complex urban challenge and a basic understanding of the amount and composition of materials generated is significantly limited by the inability to directly measure waste disposal as a result of disaggregated collection systems and rapidly changing waste streams caused by a variety of socio-economic factors [11, 72]. Furthermore, specific data sources for MSW generation are often unavailable or aggregated to regional or administrative scales preventing granular and robust statistical analysis. As such, modeling and forecasting MSW generation are important elements necessary for the planning of waste management systems and programs, including waste reduction programs, collection strategies and facilities placement, and effective and economic operations management and allocation of resources.

A variety of modeling methodologies have been used to forecast waste generation including the use of group comparison, correlation analysis, multiple regression analysis, input-output analysis, time-series analysis, and system dynamics modeling [11]. These models often focus on identifying the underlying relationship between variables that drive waste generation. For example, at the municipal level, [72] identified urban morphology, tourism activity, level of education, and income as the most influencing factors leading to MSW generation while [26] used single regression analysis to link gross domestic product and related total consumer expenditure as strong correlating factors in waste generation at the country level. [69] and [77] used traditional time-series approaches such as Autoregressive and Integrated Moving Average (ARIMA) and seasonal Autoregressive and Integrated Moving Average (sARIMA) to predict generation. [93] disregarded demographic and socioeconomic factors and forecast waste generation using a hybrid sARIMA model and grey system theory, a methodology to reveal the dynamic relationships in a system using differential equations that is derived from control theory in which the term grey describes

the understanding of information in the system (a system is defined as “grey” if the information about the system is only partially known). [94] used artificial neural networks to predict weekly waste generation in Tehran, while [1] used a combination of partial least square for feature selection and support vector machine modeling to predict for the same area.

2.3 User-Generated Data for Education and Landfill Monitoring

In January 2016, one of India’s largest landfills caught fire filling the city of Mumbai with thick smoke that lasted for several weeks and was visible from space [86]. Thousands of people, many of the poorest in Mumbai, were exposed to the hazardous smoke daily. In 2015, a landfill in Shenzhen, China collapsed resulting in a landslide killing 69 people [14]. Similar landfill-related deaths have taken place in the Philippines in 2000, Athens (Greece) in 2003 and Bandung (Indonesia) in 2005 [53]. Investigations into the causes of these events continually reveal evidence of mismanagement, inadequate oversight and insufficient regulations. In the United States (US), landfills continue to play an important role in waste disposal. In 2012, 251 million tons of MSW was generated, of which 53% was disposed in landfills [88]. Though the percentage of solid waste disposed in landfills has decreased in the United States since the 1980’s due to increased recovery rates, overall MSW generation has increased and landfills continue to be the primary method of waste disposal due to their simplicity and cost-effectiveness [30, 8].

The potentially dangerous, yet unshakable role of landfills in society reveals an intricate and complex relationship that spans numerous areas including politics, economics and environmental justice. Citizens’ mentality towards waste, and its consequences, is often best described as ‘out of sight, out of mind’. For most individuals, waste is collected curbside and never seen again. While this context describes most of society, significant portions of the population experi-

ence the effects of landfills daily. [91] have described the social justice impact of landfills in the US by highlighting the disproportionate share of landfills being built and operated in low-income and working class communities, as well as communities of color. For these communities, waste is not whisked away to some remote location, but deposited in their own backyard. The effect of this ‘out of sight, out of mind’ mentality, combined with disjointed systems of waste collection and the undesirable nature of landfills, is a waste management system that is abstracted from much of society and, in particular, lacks transparency concerning its administration, operation and maintenance. Access to information, greater transparency and increased public awareness about the role of landfills in society are therefore essential elements necessary to mitigate the potential negative impacts of landfills.

The role of public participation in waste management has historically been limited to decision-making processes in which local governments seek public input to determine the placement of waste management infrastructure [17]. Recently, however, community-based environmental monitoring efforts have proven to be an important aspect of public participation that can provide government agencies tacit knowledge about local ecosystems [24]. In El Rama, Nicaragua, for example, members of the Public Laboratory for Open Technology and Science used balloon photography to create geo-located maps of a nearby landfill [25]. This data collection effort brought together members of the local government and the informal recycling community who used these maps to evaluate the landfill’s size, material content and overall impact on the surrounding wetlands in order to collectively identify economic opportunities for the informal recycling community.

A similar participatory data collection effort was organized in the Czech Republic to map illegal dumping or overflowing waste sites [57]. Researchers facilitated citizen participation by developing a smartphone/mobile application to identify and report areas used for illegal dumping. The crowdsourced effort identified 1,438 illegal dumping sites, and as a result, 200 sites were cleaned

by the city. This effort demonstrates not only the potential for the public to participate in identifying problem areas, but also the ability for citizen generated data to inform the city’s cleanup efforts.

The Spanish-based non-profit organization, Basurama, is an artist collective that incorporates public participation in the creation of artwork and performances that explore culture and the environment [65]. Basurama, meaning love trash, uses the theme of waste to develop new perspectives and attitudes about the processes of consumption and waste generation. The organization collaborates with local community organizations to create participatory learning opportunities including the creation of interactive large-scale public art displays, developing and launching re-use initiatives and organizing workshops and visits to document local landfills.

2.4 Sensor Data for Monitoring Air Quality

Air quality is an important quality of life concern with well-established links to serious respiratory illnesses, cardiovascular disease, and increased mortality rates [75]. Cities in particular often experience high levels of fine particulate matter (PM_{2.5}), especially in developing countries where industrial expansion has created unprecedented levels of poor air quality [20]. In order to measure and evaluate levels of PM_{2.5}, government agencies often operate air quality monitoring stations that provide ambient PM_{2.5} concentration measurements. In London, for example, researchers at King’s College London partnered with local authorities to create a network of approximately 200 air quality monitors that collect and measure a range of pollutants throughout the city [62]. Similarly, the New York State Department of Environmental Conservation (DEC) operates a network of approximately 20 air quality monitors in NYC that report hourly measurements for a variety of airborne pollutants. While these networks may provide the necessary information for regulatory purposes, they often fail to capture the granular spatiotemporal variations in PM_{2.5} levels that can occur

over short distances ($<1\text{km}$) [18]. Urban environments in particular, contain widely varying mixing ratios with diverse and complex emission sources that require high resolution spatial and temporal monitoring networks to adequately quantify and describe air quality [66].

The proliferation of low-cost sensor technologies offers new opportunities to monitor and study air quality in urban environments. A growing body of research has begun to use low-cost aerosol monitors to provide high resolution spatiotemporal measurements by creating dense spatial networks that can inform local and regional emission sources' contribution to total pollution levels, as well as increase the ability to identify pollution hot-spots [40, 46, 81, 63, 67]. Furthermore, these Wi-Fi enabled devices are often compact, low-powered and easy to operate offering the ability to establish and facilitate participatory networks[48, 82]. Dense air quality monitoring networks also enable community-based feedback loops that can be used to both protect individuals susceptible to poor air quality and identify specific causes of particulate matter pollution.

While low-cost devices offer new opportunities to increase the resolution of air quality measurements, there are several important limitations to be considered. The central challenge of using low-cost devices is ensuring data quality [82, 59]. Though air quality monitoring programs implemented by federal, state and local authorities require significantly higher costs, they also operate under standard procedures for calibration, data collection, and data post-processing methods, which ensure consistency across devices. In contrast, low-cost devices often suffer from a lack of manufacturer information about the specific operation and limitations of the device, as well as employ simplistic sampling techniques that fundamentally inhibit the device's performance ability. Furthermore, low-cost sensors often require individual and frequent calibration, which involves regular access to expensive equipment and expertise, and can be impractical for large-scale deployments.

2.5 Chapter Summary

This chapter has provided important background information necessary to lay the foundation and rationale for the specific studies presented in this thesis. The chapter defines data in a urban context and presents a taxonomy of urban data, which is formulated by describing how the underlying source data are generated. This taxonomy broadly classifies urban data into categories of administrative data, user-generated data and sensor data, and provides a definition for each data type along with specific examples and background information. While an extensive taxonomy of urban data is beyond the scope of this thesis, the taxonomy presented here does provide the foundation for exploring the usefulness of urban data and identifying specific characteristics of urban data that can be generalized and evaluated.

To evaluate the usefulness of urban data, the studies presented in this thesis aim to use a specific urban data type derived from the taxonomy to explore and study a complex and dynamic urban phenomena. Specifically, these studies focus on the phenomena of waste management and air quality, which are not only complex socio-environmental processes well-suited for urban informatics studies, but also fundamental urban challenges towards creating sustainable cities. As such, this chapter continues from a taxonomy of urban data to specific background research related to the individual studies presented in Chapters 3-5. This background research covers work related to modeling waste generation using administrative data, user-generated data obtained through crowdsourcing initiatives and using sensor data generated from low-cost hardware to monitor air quality.

CHAPTER 3

Investigating Urban Data: A Gradient Boosting Model for Short-term Waste Prediction in New York City

Waste management continues to be one of the most challenging issues facing cities today. While the generation of residential waste has continued for centuries, recent technological advances now offer new opportunities to collect information about waste generation that can significantly improve efforts to manage and provide collection services. New York City in particular, who faces the tremendous challenge of collecting waste from its nearly 8.4 million inhabitants, has been collecting information about waste generation for over a decade. The information collected by the New York City Department of Sanitation (DSNY) describes with unprecedented detail the amount of waste collected, the type of trash or recycling collected, and the time and location of collection for the entire city. This administrative dataset is exemplary of the new sources of information motivating the field of urban informatics and the work in this chapter aims to evaluate the potential of such data.

The goal of the research presented in this chapter is to use historical municipal solid waste (MSW) data supplied by DSNY to forecast waste generation across the city. While the supplied dataset offers numerous possibilities to study other waste-related phenomena, the decision was made, together with DSNY officials, to develop a predictive model in order to identify new opportunities for DSNY to improve resource allocation and lay the foundation for future work desired by the agency. Indeed, the work presented in this chapter led to a follow up study focused on predicting waste generation at the building-level [54]. The study was not included in this thesis since it was largely based on the same administrative dataset and did not advance the insights into the usefulness of administrative

data that were derived from the study presented in this chapter.

This work also builds upon existing data-centric forecasting approaches in several ways. First, the close collaboration with the DSNY provided detailed information about citywide operations and the specific challenges faced by the agency. The agency’s tacit knowledge of the city’s waste system was important for proper data organization and cleaning processes, including specific information about how source data was generated, as well as provided insight on and confirmation of preliminary analyses. Second, the breadth and depth of the historical data provided by the DSNY is unprecedented in urban waste forecasting studies. Not only is this dataset highly granular both temporally and spatially, it also spans a full ten-years that allows for robust statistical results and thorough model cross-validation. Finally, this research uniquely uses a Gradient Boosting Regression model for forecasting in both time and space for New York City.

3.1 Waste in NYC

Currently all of New York City (NYC)’s waste is exported out of the city through a network of contract vendors. These vendors use a combination of long-haul trailer trucks (48%), trains (42%), and direct haul (10%). At present, 80% of the city’s solid waste is disposed of in landfills located in New York, Pennsylvania, Ohio, South Carolina, Virginia, and Kentucky, and 20% is disposed of in waste-to-energy facilities in New York, New Jersey, Pennsylvania, and Connecticut. The operational budget for the DSNY in 2012 was \$1.6 billion dollars [50].

In 2007, NYC Mayor Michael Bloomberg launched PlaNYC that established the goal of diverting 75% of the city’s solid waste from landfills by 2030 [13]. To achieve this goal, the DSNY established a pilot organic collection program to capture food scraps, yard clippings and soiled paper that serviced 100,000 homes and 40% of NYC schools as well as enhanced drop-off programs for diverting other waste including textiles, e-waste and household hazardous materials. The

recycling program was expanded in 2013 as a result of the construction of a new state-of-the art facility which allowed the collection of all rigid plastics for recycling.

In 2015, Mayor de Blasio announced the OneNYC plan that, among other initiatives, sets a citywide goal of 90% reduction of waste disposed in landfills by 2030 [27]. To achieve this goal, the city aims to expand its organics collection program to the entire city, create a single-stream recycling program to enhance curb-side collection, expand the recycling program to include New York City Housing Authority buildings, reduce the use of non-compostable wastes (plastic bags) and initiate zero waste programs in NYC schools.

3.2 DSNY data

To manage the waste generated by NYC's ~ 8.4 million inhabitants, DSNY employs over 7,000 sanitation workers, servicing 59 community districts. In total, 47,000 tons of municipal solid waste is produced in the city each day. The agency's purview includes collection of waste from city residents, public agencies and nonprofit organizations as well as street cleaning and snow clearing mandates. DSNY collects $\sim 25\%$ of the total waste produced in NYC. The remaining 75% is handled by private haulers and includes commercial/business waste, construction and demolition waste, and industrial waste.

DSNY provides bi-weekly or tri-weekly MSW collection services throughout the city as well as recycling collection once per week. NYC residents are required to separate recyclables into two separate bins, one comprised of metal, glass and plastic (MGP) and the other with mixed cardboard and paper. The city's recycling program began in 1989 though it was suspended for two years from 2002 to 2004.

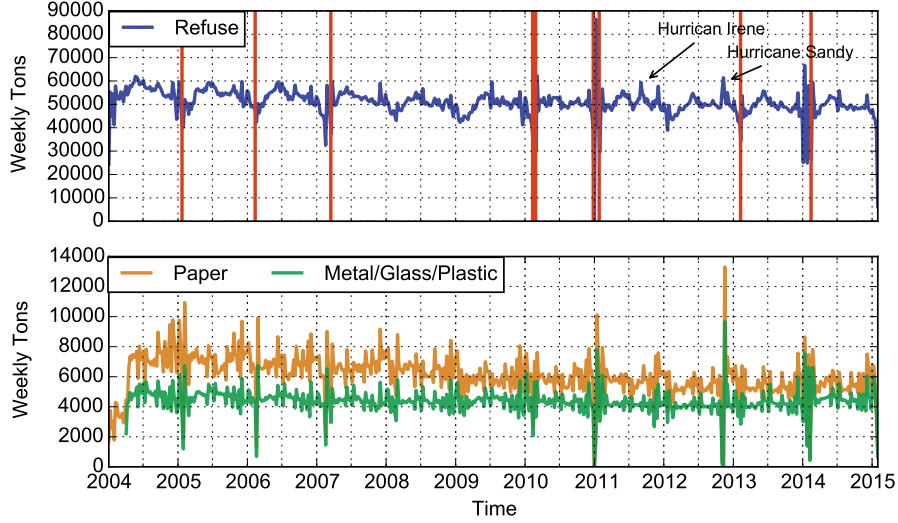


Figure 3.1: New York City’s weekly refuse and recycling collection tonnages from 2004-2015. The vertical red lines indicate significant winter weather events with above average snowfall.

3.3 Data Collection

The administrative data provided by DSNY spans more than a decade, from July 2003 to January 2015. Each record in this dataset contains the collection information for a single truck. Specifically, each record holds a unique truck ID, the collected tonnage inferred by weighing the truck, the time the truck was weighed, the type of material collected (refuse, MGP, paper), and the geospatial area from which the waste was collected (DSNY uses 232 geographies across NYC’s five boroughs that are referred to as sections).

Figure 3.1 shows the total weekly collection tonnage integrated across all sections of the city for both refuse and recycling. There are clear temporal patterns at multiple timescales. For example, strong seasonality is apparent with higher waste generation rates during the summer and lower generation rates during the winter. This observation is consistent with previous research [56, 28]. A decreasing trend in refuse generation can also be identified: in 2005, the average weekly refuse generation was approximately 60,000 tons, which has slowly declined to 48,000 tons per week in 2014 despite an increasing urban

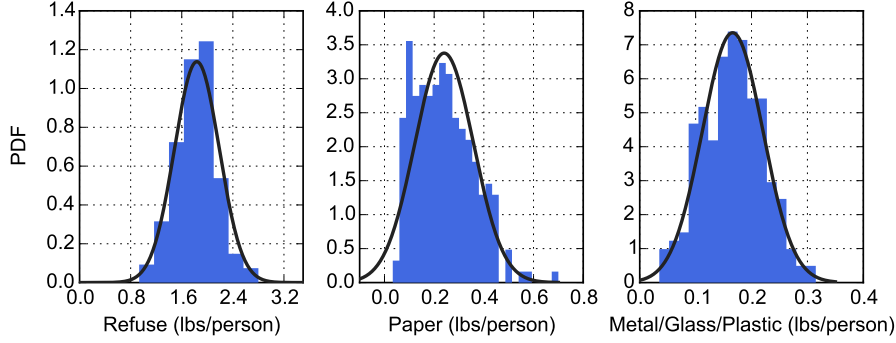


Figure 3.2: Daily per capita distribution of waste and recycling for DSNY sections. The black lines indicate a normal distribution.

population. This overall decline in waste generation is in part due to the effects of the recession in 2008 as well as reduced products and product packaging [37]. NYC’s recycling program initiated in 2004 and has maintained an average recycling rate of 17% .

In January of 2011, a winter storm produced 20 inches of snow resulting in a significant aberration in waste collection, visible by the sharp increase and decrease in tonnage. The snowfall paralyzed the city’s means of transportation and left vehicles abandoned roadside reducing DSNY’s ability to clear snow. Subsequently, waste collection was delayed across the city. This interruption is visible in the sharp decline in tonnage during the event and an increase in tonnage following the event when DSNY began collecting refuse and recycling that accumulated during snow clearing operations.

Recycling rates vary across DSNY collection sections. Figure 3.2 shows the distribution of per capita waste generation that highlights a narrow distribution for the refuse stream ranging from 1 pound to 2.5 pounds, while the paper and MGP recycling streams show a much larger distribution of per capita daily generation. [22] suggest that the variance in recycling rates in NYC are strongly correlated to four variables including percentage of persons below the poverty level, percentage of households headed by a single female with children, percentage of adults without a high school diploma and the percent of minority

population.

3.4 Methods

Figures 1-2 demonstrate that there is tremendous complexity in the data gathered by DSNY, though there are also clear patterns that can be extracted for the purpose of making near-term (~ 1 -2 weeks in advance) spatiotemporal predictions of tonnages. As noted above, such predictions would be useful for both day-to-day operations and long term planning. We now present a machine learning (ML) model designed to make such predictions using information from the data stream itself as well as external temporal data sets.

3.4.1 Model Building

A Gradient Boosting Regression Tree (GBRT) model was chosen to predict short-term waste generation. GBRT is based on decision tree regression models that are often used in ML because of their ability to handle non-linear and complex relationships between data. A typical decision tree model uses binary splitting to divide a feature space into regions and fit a different (linear) model to each region. This process is performed recursively and at each stage, the split point is determined by the greatest reduction in residual sum of squares (RSS). The process results in a single tree-like structure that “best” describes the underlying relationship between variables in a dataset [31].

GBRT extends traditional decision tree modeling by incorporating a statistical technique called boosting. Instead of building one “best fit” tree model, boosting improves prediction by building many “weak” models that are then aggregated to form a single consensus model. Weak models are defined as performing only slightly better than random guessing. Trees are created sequentially using the residuals of the first tree as the input of the new tree. In this manner, the model learns the relationship between features by continually re-weighting the input data based on the errors of the previous tree; features

performing poorly receive an increased weight [34].

Several parameters are tuned to optimize model performance. The number of trees and their depths control the final tree’s structure and complexity. Tree depth is particularly significant because it determines the degree of interaction between features. Since trees are grown sequentially, each new tree takes into account the previous trees and therefore shallow trees with a depth of 4-6 are often preferred. The model’s learning rate is another important parameter that determines how much each tree contributes to the overall model. A low learning rate will increase the number of trees used, which is ideal for better performance, but requires an increase in computation time [74].

GBRT has several advantages including the ability to process heterogeneous data, the support of different loss functions and the ability to automatically determine non-linear relationships between feature interactions [44].

Two GBRT models were built. A spatiotemporal model was built whose features include spatial and temporal characteristics, while a second, time-integrated spatial model was built in order to validate the spatial relationships between features. The spatiotemporal model was trained using weekly data from 2005 to 2011 and weekly predictions were made for each of the 232 DSNY sections for the year 2012. Validation was performed following each weekly prediction and the coefficient of determination (R^2) scores were averaged over the 52 prediction weeks. After each prediction iteration, the actual tonnage from the predicted week was then incorporated into the training set for the next iteration. The number of trees built for each model was chosen to be 1000. The model’s learning rate was set to 0.1 and the maximum number of features in each tree was 6.

The spatial model used the same information as the spatiotemporal model but was trained and tested differently. Instead of dividing the training data based on the temporal aspects of the data (i.e. 2005-2011), the spatial model divided the training data based on the spatial characteristics (i.e. DSNY sections) of the data, which was accomplished by training the model on all data

from 192 randomly selected sections ($\sim 83\%$) and predicting for the remaining 40 sections. 10-fold cross-validation was performed and results were averaged. Model parameters were identical to the temporal model.

Metrics for model performance were calculated using root mean squared error (RMSE), root mean squared logarithmic error (RMSLE) and R^2 . RMSE is a common metric to measure prediction performance while RMSLE penalizes an under-predicted estimate greater than an over-predicted estimate. This quality makes it suitable for the DSNY operations assuming over prediction is preferable compared to under-prediction, from an agency service standpoint. RMSLE is calculated as

$$RMSLE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\log(p_i + 1) - \log(a_i + 1))^2}$$

where p_i is the predicted value and a_i is the actual observation.

The working models include 28 features and predictions were made across three different waste streams including refuse, paper recycling and MGP recycling. Each model was also run three separate times with different feature groups. The first iteration used all features while the second iteration used only internal features and third iteration used only external features (see Section 3.4.2). For the prediction week, weather averages, holidays and disaster notifications are assumed from weekly forecasts.

3.4.2 Feature Engineering

Feature selection resulted in the development of two groups of features; internal and external features. Internal features are generated from the DSNY waste dataset while external features are generated from alternate data sources. Features were selected either because of their strong relationship to waste generation as indicated through previous research or because of their overall temporal and/or spatial granularity and significance in regards to the study period and area.

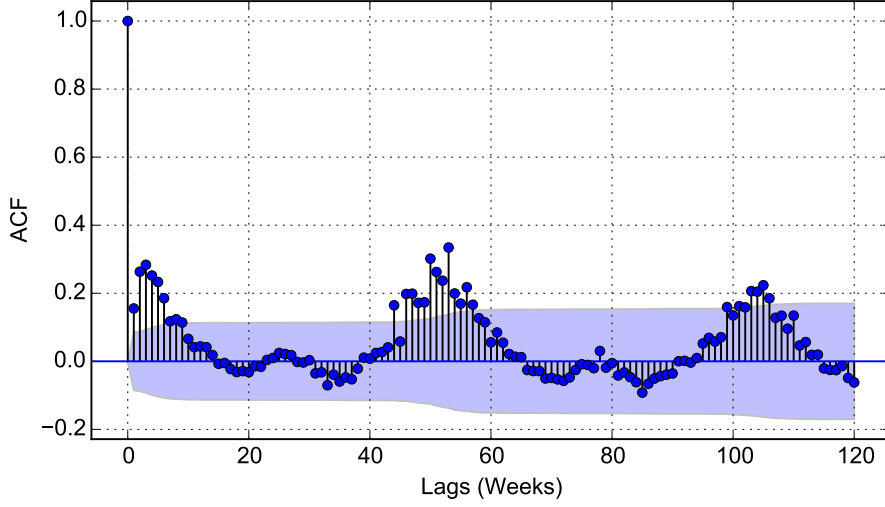


Figure 3.3: Autocorrelation with 95 confidence intervals shown in light blue where the standard deviation is computed using Barlett’s formula.

Figure 3.3 shows the autocorrelation of tonnage as a function of time for lags between 0 and 120 weeks. The significant peaks around 52 (and 104 weeks, i.e., annual correlation) and 4 weeks suggest several internal features for our GBRT model: the previous year’s tonnage, the tonnage four weeks prior, and the previous week’s tonnage.

External features are generated from outside of the supplied DSNY dataset. Demographic and socioeconomic data were obtained from the American Community Survey (ACS) and Longitudinal Employer-Household Dynamics (LEHD) data that include total population, three age brackets, three income brackets, five race groups, and four categories of educational attainment. Other external features were included for their significant temporal and/or spatial granularity. These data include historical weather and data derived from the New York City’s Department of Finance Primary Land Use and Tax Lot Output (PLUTO) data. Features included from PLUTO were the average lot market value, the total number of residential units, population density (population divided by the total residential area), and the percent of residential lots.

Information about National American holidays was also included due to the

3. Investigating Urban Data: A Gradient Boosting Model for Short-term Waste Prediction in New York City

Data Source	Model Features
DSNY	Week minus 52 weeks Week minus 4 weeks Week minus 1 weeks
PLUTO	Residential landuse percentage* Number of residential units Population density** Average total lot value
Weather Underground	Average weekly temperature Total weekly precipitation Average weekly humidity Weather events***
Data.gov	National Holidays
LEHD	Age 29 or younger Age 30 to 54 Age 55 or older Earning \$1,250/month or less Earning \$1251/month to \$3,333/month Earning greater than \$3,333/month Race: White, Alone Race: Black or African American Alone Race: American Indian or Alaska Native Alone Race: Asian Alone Race: Native Hawaiian or Other Pacific Islander Alone Educational Attainment: Less than high school Educational Level: High school equivalent, no college Educational Level: Some college or Associate Degree Educational Level: Bachelor degree or advanced degree

* Landuse categories 01, 02, 03, 04

** Section population divided by the sum of the section residential area

*** Binary sum of hourly weather observations (rain, snow, fog etc..)

Table 3.1: Complete list of features used for prediction

noticeable peaks following July 4th, December 25th etc. Table 3.1 provides a complete list of features.

Many considerations were taken in aligning these datasets given the diverse temporal and spatial nature of the data. Temporally, in order to appropriately align the multiple datasets, only data from years 2005 through 2012 were used. The supplied DSNY data was also aggregated to the week level generating a total weekly tonnage for individual geographies across the city. The DSNY does not provide daily collection for each household and at most, collection occurs tri-weekly. Therefore, the total weekly collection tonnage was chosen as a common temporal scale in order to compare different sections across the city. This aggregation also mitigates the numerous fluctuations and variations that occur within individual sections and makes waste generation patterns apparent at the section level. Weather data included average weekly temperature, average weekly humidity, the weekly precipitation total and the number of days with weather events.

Spatial data aggregation was also necessary in some instances. For example, both the PLUTO dataset and the LEHD dataset exist at a more granular spatial resolution than the DSNY waste data. In these cases, data were joined spatially to match the DSNY spatial dimensions and the sum or mean value was used. Figure 3.4 is a map of Lower Manhattan and showcases the spatial dimensions of a typical DSNY section and the census and PLUTO sub-units.

3.5 Results and Discussion

3.5.1 Model Performance

Table 3.2 shows prediction results for both the spatiotemporal model and the spatial model for each waste stream. Overall spatiotemporal model performance shows good prediction accuracy for all waste streams, though performance varies per stream and per feature group (i.e. internal features only, external features only, all features). The results show that prediction for the refuse stream con-

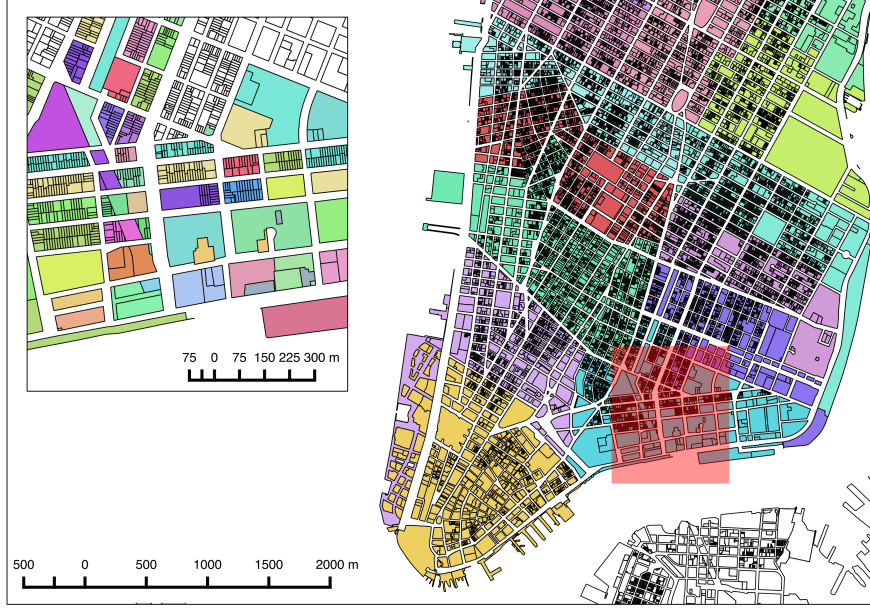


Figure 3.4: DSNY waste collection map. Large multi-colored areas indicate a DSNY section while smaller multi-colored areas reflect census blocks.

sistently performs better than the other two waste streams with approximately 15% increase in accuracy over paper predictions and a 20% increase over MGP predictions. When using all features, the average refuse prediction accuracy is 88% while the average prediction accuracy for paper and MGP streams are 74% and 68% respectively. Spatiotemporal model performance using only internal features performs better than using only external features for each waste stream. Similar to the spatiotemporal model, the results for the spatial model also show good prediction accuracy for refuse, though decrease for the recycling streams.

Figure 3.5 shows refuse prediction results for the spatiotemporal model using multiple feature groups. The weekly tonnage results are aggregated from individual predictions at the DSNY section level to demonstrate waste generation predictions city-wide. Figure 3.6 shows refuse prediction results for an individual section in Manhattan using the spatiotemporal model.

3. Investigating Urban Data: A Gradient Boosting Model for Short-term Waste Prediction in New York City

Stream	Group	R2		RMSLE	
		Spatial	Spatiotemporal	Spatial	Spatiotemporal
Refuse	All features	0.906	0.889	0.220	0.105
	External	0.604	0.837	0.288	0.145
	Internal	0.871	0.875	0.283	0.108
Paper	All features	0.791	0.744	0.374	0.313
	External	0.628	0.508	0.439	0.381
	Internal	0.738	0.731	0.428	0.314
MGP	All features	0.694	0.685	0.358	0.307
	External	0.428	0.578	0.413	0.354
	Internal	0.606	0.658	0.410	0.315

Table 3.2: Model prediction results

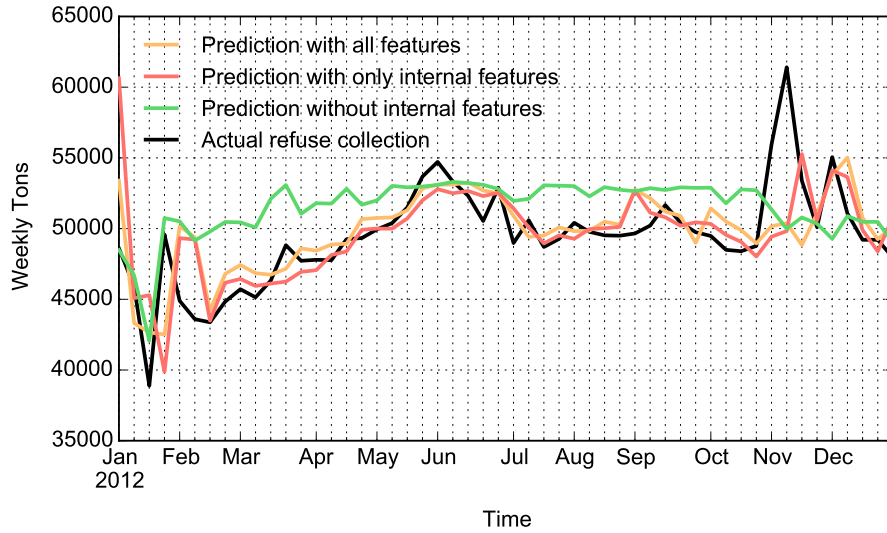


Figure 3.5: Spatiotemporal refuse prediction integrated for all sections using multiple feature groups.

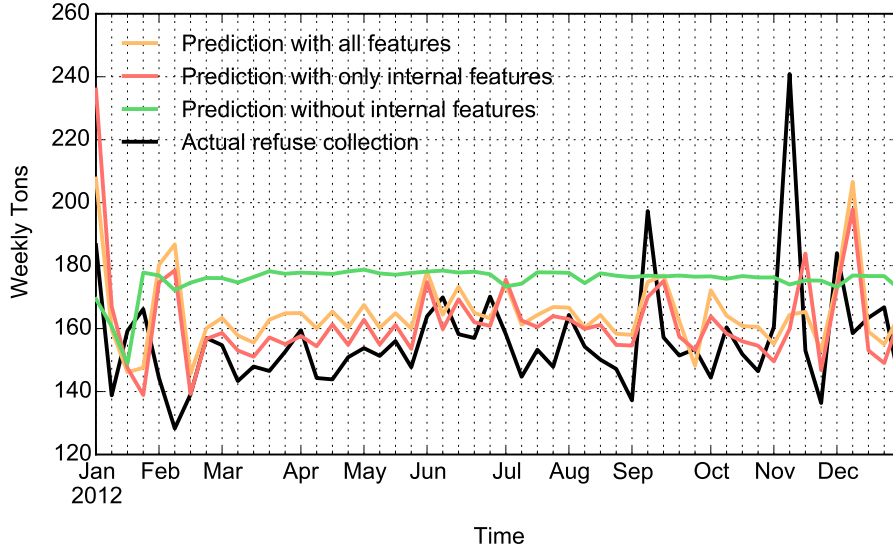


Figure 3.6: Spatiotemporal refuse prediction for an individual section in Manhattan.

3.5.2 Feature Importance

Figure 3.7 highlights the top 20 features ranked by their relative feature importance for refuse prediction using the spatiotemporal and spatial models. Scores were averaged over all model iterations. Relative feature importance is an indicator of a feature’s contribution in predicting a target response and is determined by how often a feature is used in the split points of a tree. The more often a feature is used, the greater the feature’s importance. For ensemble tree models, values are averaged for each feature and the sum of all feature importances is equal to 1.

The relative feature importance for both the spatiotemporal model and the spatial model indicates the same top ten features are the most significant features used in the models and comprise approximately 90% of the relative feature importance. Similarly, the same top five features collectively account for approximately 75% of the relative feature importance. All three internal features rank amongst the top five features, though the most significant feature used in the model is temperature. All weather features are included in the top ten relative

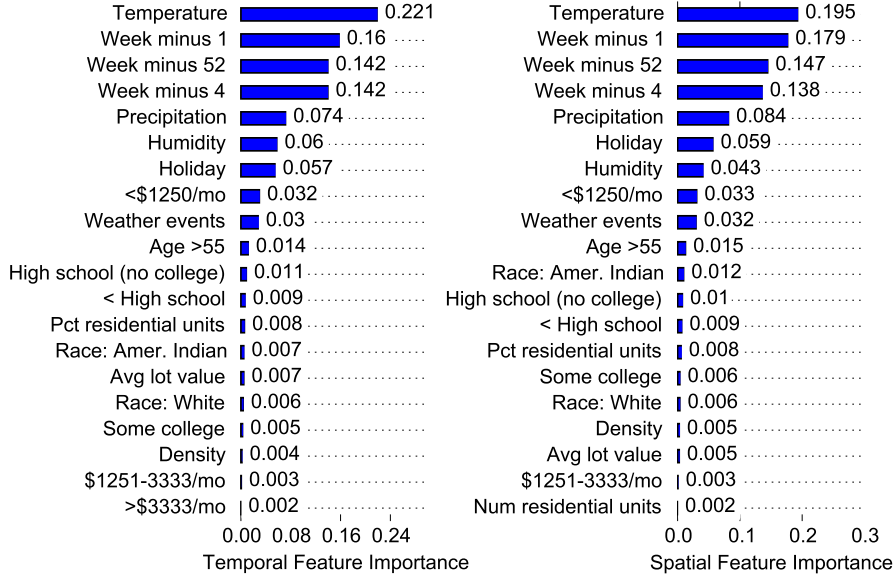


Figure 3.7: Importance ranking for internal and external features used in refuse prediction.

feature importance.

3.5.3 Detailed Discussion of Findings

Overall, the best model performance occurs when predicting refuse compared to predicting either recycling stream. We conclude this results from the selected features used in the model, granular spatial and temporal data for model training, and most importantly, the regular nature of refuse generated in NYC. This regularity is observed in both space and time and persists through disruptions in collection caused by extreme events. In order to avoid a buildup of uncollected waste, every under-collection must result in a corresponding “make-up” collection following shortly thereafter. For example, this can be seen in Figure 3.1 where sharp declines in collection tonnages that can be explained by exceptional events (e.g., winter storms) that disrupted DSNY collection efforts are immediately followed by an increase in collection following the event. This demonstrates that while the DSNY collection efforts were interrupted, public waste generation continued and strongly supports the argument that the data

aggregated to the week more accurately represents the regularity and steady flow of waste generated by society.

Figures 3.5-3.6 show highly accurate prediction results throughout the year with the exception of two extreme weather events including an above average snowfall in late January and Hurricane Sandy in late November. The nature of these events make them extremely difficult to predict and overall prediction accuracy decreases during these events. However, by including weather information, the model does have the ability to rebound from these events.

Another important observation relating to extreme weather is the impact of historical weather on current predictions. Neither model directly includes historical weather data beyond historical data used for training purposes. Each prediction is made considering only the weather for the prediction week. However, each prediction does include the previous year's tonnage, which in some cases, is affected by previous extreme weather events. For example, as previously mentioned an above average snowfall took place in late January of 2012. After investigation, it was observed that a similar extreme winter event took place the previous week exactly one year earlier. From the model's perspective, the previous storm impacted the previous year's tonnage, which is a feature included for prediction and can therefore significantly alter prediction results. These fluctuations can be observed in Figure 3.5 during the last two weeks of January and the beginning of February.

Feature importance results provide insight into the model's decision-making process but should not be interpreted as causality or the strength of the relationship between the dependent and independent variables. For both spatial and spatiotemporal models, temperature is ranked as the most significant feature. One interpretation is that temperature contains the high frequency time-scale information which captures weekly fluctuations as well as provides seasonal information capturing the low frequency patterns in refuse generation as previously observed. However, a comparison of models using internal features versus external features clearly shows that models using internal features

perform better than models with external features only. One can conclude that while weather is an important element for prediction providing high and low frequency information, the inclusion of waste data yields a better model for prediction.

Another important observation is the higher R^2 scores for spatial models compared to spatiotemporal models. The high R^2 scores for the spatial model validates the spatial features and indicates that space is an important element for prediction. If the spatial features were insignificant, one would expect poor model performance. The superior performance of the spatial model compared to the spatiotemporal model suggests that removing temporal information from the waste data yields a more accurate predictive model. However, while the spatial model does produce a higher R^2 score, it is important to consider the RMSLE results of each model. Indeed, the RMSLE score for the spatiotemporal model consistently predicts with less error highlighting a sensitivity between the two models that makes direct comparison difficult. One explanation for this is the methodology used for each model in which the spatial model predicts waste for 40 sections over 416 weeks for each iteration, while the spatiotemporal model predicts 232 sections for one week each iteration. Because results are averaged, the spatiotemporal model is therefore more sensitive to large errors often resulting from extreme weather events.

Similarly, it should also be noted that the high scores for the spatial model might also be inflated because of spatial relationships that exist between independent variables. The existence of spatial clusters, if present, will likely be affected by the aggregation of independent variables to larger spatial units, and spatial correlations between these spatial units may adversely affect model results. Future work to identify possible spatial autocorrelation should be considered in order to fully identify and understand the spatial relationships between independent variables and their effect on the overall predictive model.

The accuracy of the refuse model can also be translated directly to DSNY operations. The spatiotemporal refuse model produces an RMSE score of 31.39

and a mean absolute error (MAE) of 14.68. Because RMSE and MAE have the same unit of measurement as the dependent variable, results can be interpreted as individual section predictions are on average within ± 32 tons of the reported tonnage. This is significant given that the capacity of a DSNY truck is approximately 12 tons, which suggests that re-allocation of DSNY trucks for increased efficiency results in a ± 3 truck accuracy.

Lastly, we note that the high granularity in space and time of the DSNY data provides the most significant predictive power needed for forecasting waste generation (and indeed, our results show that historical waste data alone is sufficient in all but the most extreme cases). With only one exception, the absence of waste data consistently results in poor model performance. The effort and ability of the DSNY to collect data about waste collection serves as a strong example for other cities aiming to improve the understanding and operations through data-driven analysis. The collection and integration of data is fundamental to enable data-driven decisions for enhanced operations, improved planning and ultimately for better policy decision-making.

3.6 Discussion of Administrative Data

This research demonstrates several important advantages and challenges related to the use of administrative data in the field of urban informatics. One of the most significant observations from this study is the tremendous amount of information that can be obtained through the use of an administrative dataset. In this study, a fundamental descriptive analysis of data provided by DSNY resulted in important insights that not only informed officials about citywide patterns including seasonal trends in waste generation and participation in recycling programs, but also provided the foundational information used to develop a predictive model. Furthermore, the temporal and spatial detail of the source data enabled robust statistical tests and rigorous model validation techniques, and demonstrated that a single administrative dataset can provide substantial

information for successfully modeling and simulating urban phenomena.

While the ability of this stand-alone dataset is substantial, it should also be noted that in order to use the dataset to its full potential, integration with other administrative data was necessary. As demonstrated in this work, the addition of other administrative datasets, such as weather, demographic and socio-economic data, enhanced the overall performance of the predictive model. Furthermore, the integration of multiple administrative datasets provided greater insight into the factors used by the model to predict. Through an analysis of feature importance, researchers can further assess the underlying factors influencing a particular urban phenomena. While the overall goal of this work was to produce a predictive model, and not necessarily an explanatory model, it is worth noting the inclusion of these other datasets allowed for greater insight into waste generation.

One of the general challenges with this research, in the broader context of urban informatics research, is the overall lack of availability of this information. While this work critically identifies and promotes the efforts of the DSNY to provide detailed waste generation information, public access to this information is limited. The only publicly available data published on NYC's Open Data platform are monthly tonnage reports for the 59 community districts. In some scenarios, this limited information may be sufficient, though based on the research presented here, this information is insufficient for any significant urban-modeling related work.

A final quality concerning administrative data is the overall usefulness in a variety of situations. As previously discussed, administrative datasets provide foundational information necessary for quantitative urban research. While the following chapters expand upon other urban data types, inherent in each of these studies is some use of administrative data. Indeed, one could argue that access to administrative data in various forms has enabled all of the research presented in this thesis. For urban informatics research, the ability to access and obtain administrative data sources is paramount.

3.7 Chapter Summary

This chapter has focused on defining and describing administrative data through the development of a short-term predictive model for waste generation in NYC. Multiple administrative datasets were integrated and combined with features from historical waste collection data by the New York City Department of Sanitation and a gradient boosting regression model was built to predict weekly waste generation across three streams with an average accuracy of 88%. The model results demonstrate that refuse generation is predictable across the city and the ability to predict waste generation can largely be explained by the regularity of waste generation by society when aggregated to one-week timescales. This regularity allowed us to derive features based purely on the waste tonnage data, which enabled excellent prediction accuracy. Importantly, the model is able to capture short timescale fluctuations associated with holidays, special events, seasonal variations and weather-related events. Furthermore, this temporal behavior is similar across spatial regions although the absolute amount of waste in different regions varies significantly. Including additional, external features (most importantly weather conditions) further improved the robustness of the model.

Broadly, this chapter outlines two important contributions. The first contribution presented in this chapter is a new method for predicting waste generation using a gradient boosting regression model. While previous research efforts have explored predicting waste generation at various temporal and spatial resolutions using a variety of traditional statistical approaches, this research uniquely implements a machine learning model capable of incorporating diverse sources of information to generate highly accurate predictions across short time spans and granular spatial areas. The application of this machine learning model also led to important insights about waste generation in cities. Specifically, the model demonstrated the ability and usefulness of historical data to forecast waste generations, as well as the usefulness of weather features. Although these weather

features do not directly influence the day-to-day generation of waste by individuals, it does serve as a proxy for seasonal patterns of waste generation, as well as the high frequency and direct effect of weather on trash that is left curbside for collection. And while the model developed in this research was not tested on other cities specifically, the underlying features contributing to the overall ability of the model are common among other cities around the world making this forecasting method adaptable and reusable by other city agencies responsible for sanitation.

The second significant contribution outlined in this chapter is a new understanding of waste generation. Specifically, this research demonstrates a regular pattern of urban waste generation with a temporal and spatial resolution not previously described. The existing knowledge about urban waste generation comes from information sources that often describe long-term trends (e.g. annual tonnages) across entire urban centers or large spatial regions. This lack of granularity has previously prevented an in-depth understanding of society's patterns of waste generation. The research presented in this chapter highlights the regular nature of waste generation as well as the effects of social, political and logistical phenomena that influence waste generation at the city level. This new insight into the phenomena of urban waste generation is not only important for future research but also for local and city agencies tasked with managing the daily stream of waste being generated.

This research was done in collaboration with the New York City Department of Sanitation in order to improve operational efficiencies for both short-term and long-term planning. The ability to forecast waste generation gives the DSNY a unique ability to optimize waste collection and vehicle allocation as well as the potential to develop targeted long-term strategies for waste reduction and recycling programs. Furthermore, this research provides a baseline for the DSNY to begin implementing and testing the effectiveness of waste reduction programs in various parts of the city.

Finally, this chapter presents an exemplary use of administrative data and

discusses some of the key advantages when using this form of data. A brief discussion of the characteristics of urban administrative data further emphasizes the usefulness and adaptability of the data, as well as the profound importance and pervasiveness of administrative data for urban informatics research.

CHAPTER 4

New Urban Data Streams: The Role of Public Participation in Urban Data

The increased availability and use of web-based and new information communication technologies (ICTs) has enabled new opportunities for public participation in urban informatics research including the potential to contribute to the processing and analysis of large-scale datasets and the ability to generate entirely new sources of information. As technology enables urban residents to become more connected, increased public participation can also serve as an important method to disseminate information and create feedback loops between citizens and government officials that can lead to mutually beneficial relationships and improve transparency, openness and efficiency. Indeed, local residents are often the first to detect disruptions in normal urban activity and can provide real-time information about neighborhood activity. While many forms of public participation exist, this chapter explores the use of online public participation to not only validate existing sources of information and generate new data sources, but also to explore the outcomes made possible through these initiatives.

As previously discussed, waste management is one of the most difficult challenges facing local government agencies for reasons that often stem from lack of public information. To address this challenge, this chapter presents a case study that employs public participation to generate information about landfills in the United States. The work describes an online crowdsourcing initiative titled “Landfill Hunter”, and explores the data-driven role played by the public in this research. Importantly, this work also provides unique insight into non-data outcomes of urban informatics research that also plays an important role advancing urban objectives.

The goal of this work is to address the often hidden nature of landfills by bridging the knowledge gap between landfills and society. Landfill Hunter is an online crowdsourcing initiative that seeks to bring the role of landfills in society to the forefront and make the presence and operation of landfills visible to the public. The initiative uses citizen participation as a pathway for learning in order to promote a greater understanding of society's consumption and waste habits, and addresses openness and transparency about landfills by providing greater access to information. To this end, Landfill Hunter uses public participation to identify active and historical landfill locations in the US and estimate their total land area.

An important aspect of Landfill Hunter is to address the lack of publicly available data about landfills. Access to information and providing openness and transparency are important aspects for education and participation. Landfill Hunter promotes openness by allowing anyone to contribute data and by using an open-source platform to enable reproducibility and further project development. The platform also provides full access to user-generated data, which is immediately and freely available. This approach creates an environment suitable for the formation of communities, both online and offline that can further motivate and reinforce learning. The data generated from this research also has several potential uses that include providing new and comprehensive insight on the size and spatial distribution of landfills in the US, the ability to compare landfills and their surroundings, and most important, the ability to monitor and track landfills over time. A unique aspect of this work lies in its ability to foster experiential learning by making visible society's waste generation habits.

4.1 Landfills in the United States

The history and evolution of landfills in society provides important insight about modern-day landfill operations and access to information. The practice of landfilling has existed for centuries, often in the form of open pit dumping. During

the 20th century, however, the dramatic change in the type and amount of waste being generated, coupled with a lack of disposal regulations, led to increasingly hazardous and unrestrained landfill practices [84]. The first federal legislation regulating landfills was the Solid Waste Management Plan followed by Resource Conservation and Recovery Act (RCRA), which defined the standards for sanitary landfilling and enabled the Environmental Protection Agency (EPA) to regulate and enforce landfilling practices [71]. Following the enactment of RCRA, the EPA began to inspect landfills across the United States (US), which led to the closure and abandonment of landfills failing to meet RCRA standards [83]. These inspections also led to the first comprehensive documentation of landfills in the US. In 1986, the EPA released the first list of all MSW landfills nationwide, which totaled 7,683 landfills. The list was updated in 1992 showing a decline in the number of landfills to 5,345, which declined even further to 3,581 in 1996, to approximately 2,300 in 1998 and 1,767 in 2002 [30, 87]. As of 2012, there are 1,908 active landfills in the US [90]. Though the total number of landfills has largely declined since the 1980s, the average size of a landfill is increasing [88].

Information about waste is often sparse and difficult to access because of disaggregated disposal methods [11]. Similarly, information about landfills is limited because they are often privately owned and operated. Indeed, very little information existed about landfills prior to RCRA because of diverse state legislation and lack of overarching federal legislation [83]. One existing source of data can be found in the EPA’s Enforcement Compliance History Online (ECHO)¹ database, which was developed as a tool for the public to view compliance and enforcement activities of over 800,000 EPA monitored facilities under the Clean Air Act, Clean Water Act and RCRA. The data are freely accessible and updated monthly [12]. Despite being the most comprehensive and centralized dataset about landfills, the database faces several challenges. In particular, the evolution of landfill regulations led to categorization gaps and abandonment of

¹<https://echo.epa.gov/>

numerous landfill sites that failed to meet RCRA standards [83]. This historical process presents fundamental challenges to collecting comprehensive data about landfills. EPA ECHO has also faced political controversy, online software glitches and the potential to be inaccessible as seen during the 2013 government shutdown [71, 35, 47].

Other sources of information about landfills can be found through studies and reports published by various institutions and non-governmental organizations. The most thorough report was a joint study by BioCycle and the Earth Engineering Center of Columbia University published in 2010 [90]. The report surveyed each individual state agency responsible for waste management to collect data about the total number of operational landfills in the US and their total remaining capacity. The data collected by the survey is based on the assumption that all states have reporting requirements and all waste management methods are reported to the appropriate state agency. The report found a total of 1,908 operational landfills, despite five states not completing the survey, and their remaining capacity ranged widely with varying units of measurement including cubic yards, years and tons.

4.2 Methodology

4.2.1 Application Design

Landfill Hunter is designed to expose participants to landfills in a novel manner. The use of Bing and Google Maps provides high-resolution maps for users to explore and virtually access landfills that are often off limits to the public (Figure 4.1). This detailed view of landfills shows participants specifically where waste is disposed, the manner in which it is disposed and the communities and environments that surround the landfill. Though landfills are often located in remote areas, a surprisingly large number of landfills are located near communities, both rural and urban. Participants witness and experience the diverse observable infrastructure including residential, commercial and industrial areas

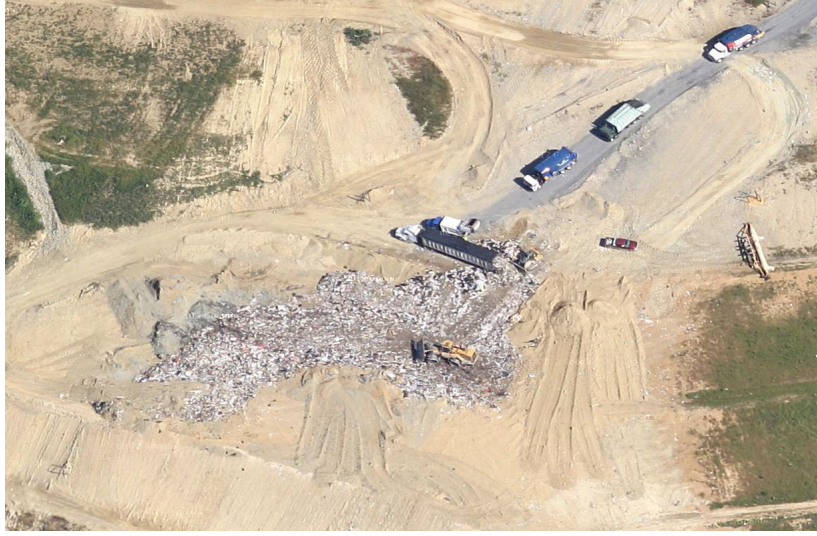


Figure 4.1: Long-haul truck dumping waste into a landfill. Screenshot taken from Google Maps.

surrounding landfills. Qualitatively, this experience often reveals surprising patterns such as the frequent placement of golf courses nearby and the placement of landfills close to bodies of water. In this case, learning about landfills takes place through the process of participation.

Landfill Hunter uses humans to navigate geospatial information and visually identify the presence of landfills. Because landfills tend to have irregular shapes and sizes, computationally identifying landfills with image analysis software and modern machine learning techniques can be difficult. Humans, however, with the proper guidance and explanation, can learn to accurately recognize, identify and map geospatial features [4, 32]. Identification tasks can furthermore be distributed to multiple users in order to provide validation and accuracy metrics that can greatly enhance data quality. The Landfill Hunter application² was designed and built with the do-it-yourself (DIY) crowdsourcing platform Crowdcrafting.org. Crowdcrafting.org is based on the open source crowdsourcing engine PyBossa, which provides the backend functionality for developing and hosting crowdsourcing applications, including generating and presenting tasks

²<http://crowdcrafting.org/project/landfill/>



Figure 4.2: Landfill images used as examples for users prior to starting. Images were taken from Google Maps.

for users to complete and the collection and storage of generated data [76].

Prior to starting, users are presented with participation instructions and objectives. These instructions explain how to navigate the application, how to outline potential landfills and how to submit answers. Most importantly, the instructions provide example landfill images to demonstrate how to visually identify a landfill. Figure 4.2 shows the six landfill images manually selected as examples to guide participants. These landfills were determined as representative landfill images because of their common features including likeness in color, barren land cover, being distinctly man-made, sharp contrast to surrounding areas and exhibit the appearance of concentric rings that indicate layering.

Upon completing the tutorial, users are presented with an online interactive map and must locate potential landfills based on previously identified coordinates from the EPA ECHO database. Using Bing maps, users can navigate and explore the area and zoom in on specific locations to examine an area in detail. If a landfill is identified, users are instructed to outline the landfill with drawing tools provided as a part of PyBossa. Users are also given the option to skip

the task if no landfill is found. To aid the user, additional information about the landfill is provided below the map including the facility name, address and state. Once a user has submitted an answer, a new task is randomly selected and presented to the user. Each task must be answered by 10 unique users before it is determined as completed.

For each task, users can draw multiple geometries, edit their entries and estimate their confidence level prior to submitting the response. After a geometry is drawn, the total area of the polygon is calculated and displayed to the user while a second calculation converts the area from meters into the equivalent number of football field lengths in order to provide the user with a more meaningful measure of the landfill's area. User-generated data is captured in GEOJSON format and results are immediately stored and made openly available by the Crowdcrafting.org backend.

Initial geospatial information about landfill locations was obtained from EPA ECHO and used as a starting point for participants to begin their landfill search. Landfills in the ECHO database were identified by using the Standard Industrial Code 4953 (Refuse Systems) and the North American Industry Classification System (NAICS) code 562212 (Solid Waste Landfill). Following the data export, further processing was necessary to remove duplicate entries, entries without geographic information and non-landfill entries. The final dataset comprised 1,235 entries representing potential landfills. Each entry contained a facility name, street address, latitude, longitude, and the number of days since last inspection. Manual verification of several potential landfill sites was performed to ensure a basic level of accuracy of the projection of the EPA ECHO data on publicly available satellite data.

4.2.2 Data Assessment and Validation

An accuracy assessment methodology was established to measure and validate data generated by Landfill Hunter. One key challenge in assessing user responses, however, is the lack of an official ground truth answer resulting from

the nature of the project; users are asked to both validate possible landfill locations and identify other potential landfills which may not exist in the presented data. Therefore, the method to assess accuracy takes an inclusive approach by rejecting only responses that appear to be entry errors and assuming all other user-generated geometries represent a landfill. Providing users the ability to skip a task supports this approach by encouraging users to answer only when they are sufficiently confident a landfill is present.

To determine entry errors for each user response, an initial test is performed to identify and remove geometries that have three or fewer points. Responses are then aggregated for each task and analyzed to identify overlapping geometries, which indicate a greater likelihood of a landfill. Similar consensus building approaches were used to identify craters on planetary surfaces finding aggregating multiple volunteer responses generated more stable results [78]. Geometries that do not overlap any other geometry and have an area that is plus or minus 5 times the standard deviation of the mean task area are considered entry errors. Using the large variation metric is intended to remove only extreme outliers that are indicative of invalid responses. Geometries that meet these two criteria are removed. The area of an individual landfill is then calculated as either the average area of the overlapping geometries or the total area of the individual non-overlapping geometry.

4.3 Results and Discussion

4.3.1 Data Contributions

To date, a total of 147 participants have completed approximately 15% of the tasks, contributing a total of 1,024 unique spatial geometries. Figure 4.3A and Figure 4.3B provide insight into how participants contributed to the project. Specifically, Figure 4.3A describes the number of responses per user, which shows only a few users contributed a majority of the responses and a large number of users contributed only a few responses. The distribution of user partici-

Table 4.1: A description of the data generated by the Landfill Hunter application.

Data Source	Name	Description
Environmental Protection Agency	Facility name	Name of facility that operates the landfill
	Facility address	Address of the operating facility
	Inspection date	Last reported inspection date by the EPA
Platform-generated	Task id	Unique identifier for the task
	Task run id	Unique identifier for platform task distribution
	Task start time	Initial time the task was presented to the user
	Task end time	Time the task was completed
	User IP Address	IP Address of the user
	User ID	Identifier if user is logged in to the Crowdcrafting platform
User-generated	Geometry	GEOJSON object with latitude and longitude points
	Confidence level	Either 'very confident', 'not confident' or 'unanswered'
	Landfill area	Land area of the landfill computed from GEOJSON object

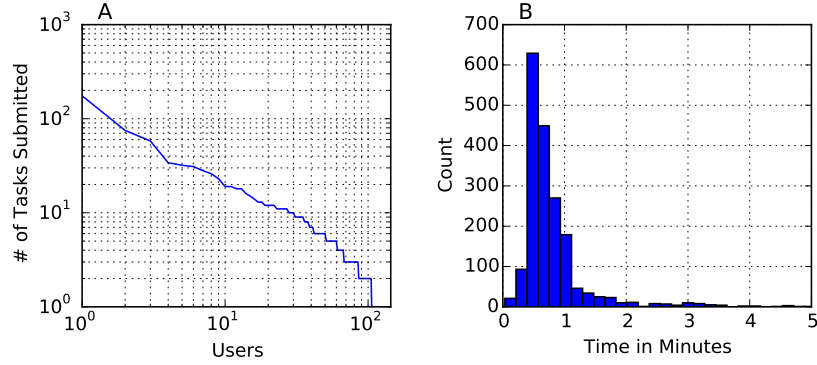


Figure 4.3: Descriptive analysis of user participation rates. Figure 4.3A describes the distribution of task responses per user and Figure 4.3B shows the distribution of time spent per task.

pation captured here is common among crowdsourcing projects often depicting variability in the level of user engagement. Figure 4.3B shows the distribution of the duration of time users spent per task. On average, users spent 55 seconds per task, which demonstrates the significant amount of engagement required in order to complete a task. While task completion time varies per project, many crowdsourcing projects require only a few seconds to complete an individual task. Similarly, users contributed 12 points on average for each task.

Data generated by the project includes a variety of information about landfills in the United States derived from both the original EPA dataset and participant responses, as well as general information about the data collection process. Table 4.1 lists the individual items included in the final dataset. A total of 729 landfills have been identified after removing outliers and combining user responses. Of the 729 user-identified landfills, 30% have been identified by multiple users as indicated by overlapping geometries.

Figure 4.4 demonstrates an example task response after aggregation and shows the diversity of potential answers. In this case, the majority of geometries overlap the same area, while three other smaller non-overlapping geometries have been identified nearby. The overlapping geometries clearly show user agreement about the location of a landfill, albeit with variations in the exact

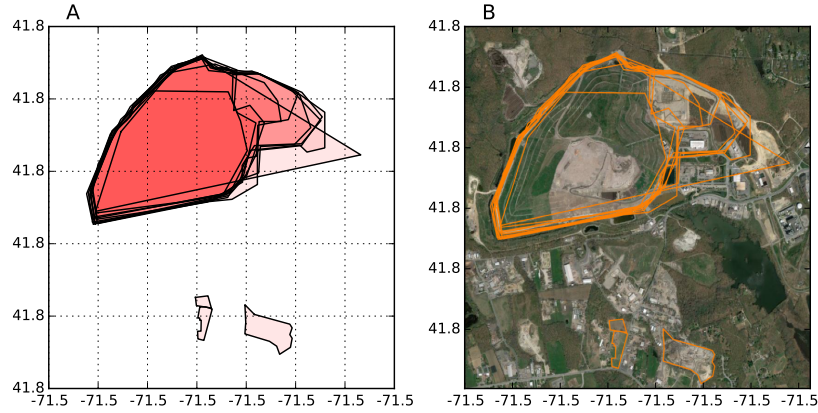


Figure 4.4: Example user responses aggregated by task. Figure 4.4A highlights user ability to identify similar landfill definitions while figure 4.4B shows the results plotted with Google Maps.

border definition. Despite being non-overlapping, the smaller geometries may also represent landfills missed by previous users. The smaller landfills are therefore also included in the results. This approach allows for users to identify potential landfills that others may have missed. Figure 4.4B shows the user generated outlines on Google Maps for verification of accuracy.

Landfill land area was aggregated per landfill and for all identified landfills. Based on user identified landfills, the average landfill area is approximately $808m^2$ and the cumulative land area of landfills in the US is $576km^2$. This estimate is based on the 729 landfills identified by users, which represents approximately 60% of the landfills from the EPA ECHO database and only 40% of the stated 1908 total landfills in the country. Using the average landfill area and the EPA's stated number of landfills, the total land area of all landfills in the US is estimated to be $1,541km^2$. This area is approximately twice the size of New York City, whose land area is $709km^2$.

4.3.2 Offline Outcomes

Participation in this crowdsourcing initiative highlights several learning opportunities both online and offline. The capacity for Landfill Hunter to create

offline collaborative experiences was demonstrated during EcoHack 2014. EcoHack was a weekend-long event that took place in New York City and brought together participants with diverse backgrounds to explore the natural environment through data and technology. The event is designed to encourage collaboration and idea sharing to extend existing research and develop new projects related to the topic of nature and technology. Selected projects were presented to participants at the start of the weekend followed by the formation of groups to begin exploring and developing new ideas over the remainder of the weekend.

A group of 15 volunteers began to work with the Landfill Hunter platform. This collaboration started by completing tasks in order to contribute data and evolved into developing similar landfill-themed sub-projects. The result of this collaboration was the development and launch of a project titled Landfill Club. The project was initially designed to encourage people to think about where their trash goes and evolved into an online space for sharing information about landfills. The aim of this online space was to develop a community by connecting residents who share the experience of living in close proximity to a landfill. The space was not only created to facilitate information sharing for the purpose of increasing awareness about the existing challenges and struggles, but also to encourage and enable these communities to share experiences, solutions and strategies towards improvement. To develop this idea, the group launched a website to collect, analyze and present data about landfills³. This site organized existing information about landfills and showcased a variety of creative landfill visualizations. For example, the group juxtaposed the geospatial data from Landfill Hunter to compare landfill shapes and sizes and to potentially identify geometric patterns in the data. Other visualizations explored mapping the distribution of landfills across the US and comparing EPA landfill inspection rates across different states.

The experience of EcoHack demonstrates the ability of a online crowdsourcing initiative to stimulate participant learning and creativity at many levels. The

³<http://landfill-club.herokuapp.com/>, <https://titanpad.com/8efVBMrPTe>

organic formation of an offline group which, extended and further explored the theme of landfills, suggests that encounters and participation in Landfill Hunter can stimulate interest and curiosity in the subject matter beyond the online experience. Participants not only identified alternative landfill challenges, but also collaboratively devised and implemented creative solutions for these challenges over the course of a weekend. The use of Landfill Hunter data for visualization and analysis furthermore demonstrates participant initiative, desire for further discovery, and a creative ability to convey information in a meaningful way.

4.3.3 Challenges and Limitations

A key challenge for this data collection initiative is the lack of a robust statistical approach for validating user responses, which resulted from limited participation. While the proposed inclusive approach for validating landfill locations and removing invalid entries provides an initial step towards validating user responses, it ultimately lacks sufficient data for evaluation. This is not only true for individual tasks that received less than the required number of responses for validation (ten responses) but also true for the entire dataset preventing the application of a cohesive and comprehensive assessment methodology.

Furthermore, while the original aim of this work was to identify all of the landfills in the United States and develop a comprehensive dataset that describes these landfills, the lack of participation suggests that alternative efforts to generate this information may be more successful at achieving this goal. However, the outcomes of this crowdsourcing initiative demonstrate benefits beyond the generation of a comprehensive dataset such as raising awareness about the presence of landfills and providing educational opportunities for individuals to better understand where their trash is ultimately disposed of. These secondary outcomes may provide sufficient value in and of themselves to justify the implementation of a crowdsourcing project using the methodology presented here. It should also be noted, that since the original focus of this research was on data collection, a full evaluation of the increase in citizen awareness and educational

outcomes was not evaluated and are presented here as speculation based on qualitative observations.

Another significant challenge relates to incentives and user motivation. Indeed, Landfill Hunter provides no extrinsic incentive to engage users beyond initial exploration of the platform and largely assumes an intrinsic participant motivation. Exacerbating this issue, significant time and effort is required of the user in order to investigate the map, identify landfills, provide an outline of the landfill and submit the results. The high level of effort demanded from the user makes task completion challenging and raises the barrier for participation [29]. To fully explore the potential of Landfill Hunter’s methodology, significant improvements to user motivation and user experience should be explored.

Several areas also exist for future work. To extend the existing initiative, critical analysis and improvement of participant motivation is needed. This includes developing appropriate incentives and rewards for participation as well as identifying and soliciting participation from specific interest groups who may directly benefit from the collection and analysis of the data. These groups may include local activist or environmental monitoring groups as well as other online based groups such as the Open Street Map community which may take interest in the project for its potential to contribute geospatial information. This type of small and targeted effort could greatly increase analytical capabilities of the project. Gamification is another possibility to enhance user motivation and has been shown to be an important factor in increasing participation and maintaining long-term engagement [43, 45].

Other possibilities to determine and collect data about landfills could also be explored. For example, other online tools such as Google Terrain view may offer opportunities to explore and identify landfills in alternative ways. Similarly, the project could be used to develop a landfill health index based on a combination of existing EPA data and citizen generated data that extends beyond geospatial information to collecting information about local communities including socioeconomic and demographic information to identify and determine patterns in

landfill activity.

4.4 Discussion of User-generated Data

This work demonstrates several unique characteristics of user-generated data obtained through the development and implementation of an online crowdsourcing initiative. In particular, this example highlights the ability for user-generated data to address difficult computational problems and generate new sources of information. One of the motivating factors for the use of public participation in this research was based on the computationally difficult task of visually identifying landfills. In this scenario, addressing a difficult urban data challenge was best overcome through distributed human computation rather than complex computational techniques. Furthermore, this work shows how user-generated data can validate existing data sources, as well as provide new sources of information in scenarios where existing sources of information are insufficient or unavailable.

While this work demonstrates participatory approaches for generating new data, numerous challenges exist related to the significant effort required to acquire these data, as well as the complexity and difficulty in processing and cleaning these data in order to obtain meaningful information. The launch of an online crowdsourcing initiative not only requires substantial work to develop the necessary infrastructure, but also requires careful consideration of the interface design and user interaction processes. Development of these crowdsourcing projects often requires an iterative testing process in order to evaluate and identify the appropriate methods for promoting proper use of the platform. Similarly, the development of a method to assess proper participation and validate user-contributed data is nontrivial. Furthermore, the success of crowdsourced initiatives fundamentally depends on soliciting a sufficient amount of participation, which requires careful consideration of the target audience and the development of appropriate outreach strategies.

Another unique characteristic of user-generated data obtained through active public participation is the ability to generate diverse and innovative outcomes at many levels. Not only can participatory research initiatives lead to insights derived from the research itself, they can also generate unexpected outcomes and foster educational and collaborative opportunities between various groups and individuals. Secondary outcomes from these initiatives such as the development of innovative solutions to similar problems or the advancement of government transparency and openness initiatives, are significant in the field of urban informatics. Educational opportunities generated by these initiatives can create positive feedback loops between participants, researchers and government officials, which can help identify problem sources and generate new solutions collectively.

Finally, the example presented in this chapter represents only one form of user-generated data, which focused on user-generated data that was obtained through active participation in a research initiative. Numerous other user-generated data collected through active processes are available including sources generated from the usage of social media platforms, offline data collection initiatives and other web-based activity, as well as other forms of passive user-generated data including the collection of call detail records or Wi-Fi probe request data generated from mobile phone usage.

4.5 Chapter Summary

This chapter has explored the usefulness of user-generated data collected through active participation in an online crowdsourcing initiative that addresses the intersection of public involvement in waste research and access to reliable information about landfills. The online data collection platform used in this work creates a unique opportunity for citizens to explore nearby and distant landfills by analyzing existing data, contributing geospatial information and exploring high-resolution maps. The outcomes of participation in this initiative include

experiential learning, generation of publicly available data and the potential to generate creative and collaborative offline experiences.

Importantly, the research conducted in this chapter has generated a new source of information about landfills. Prior to this research, public information about landfills was limited and incomplete due to the historical evolution of landfill regulation and management that left significant data gaps. While the EPA does provide substantial information about the locations of landfills, the accuracy and completeness of this data were insufficient for a comprehensive study of landfills across the country. Through the research presented in this chapter, participants in the project Landfill Hunter were able to not only validate existing sources of data about landfills, they were also able to generate entirely new information about landfills including their spatial extent, land area and specific location. The data generated by this research can significantly improve efforts to understand the impact of landfills across the country, as well as provide a novel means to monitor and evaluate these landfills over time. For example, state governments could use this information reporting and enforcement of state guidelines, as well as serve as an indicator for developing long-term waste disposal policies.

A second significant contribution of this research relates to the formation and execution of crowdsourcing research initiatives. Broadly, these initiatives have demonstrated successes in generating new sources of information and analyzing large datasets that require human participation for analysis (e.g. visual inspection of images taken of space). To accomplish these objectives, new methods for engaging and soliciting participation have been explored, especially through means of gamification, which provide users various levels of intrinsic and extrinsic motivation. However, there is very little understanding about the ability of these initiatives to generate secondary outcomes including the learning opportunities presented by participation in these projects and the secondary offline experiences that can be achieved. The research presented in this chapter provides new insight into how learning can take place through participation, and

more importantly, how collaborative experiences can be generated. While these secondary outcomes are often more difficult to quantify and evaluate, this research highlights the importance of these outcomes, especially in the context of urban informatics which can not only serve as a pathway for governments to engage citizens but also an opportunity to provide improved transparency and accountability.

Ultimately this research presents an opportunity to collect citizen-generated data about landfills and learn about the impact of landfills through participation. As explored with the subproject Landfill Club, the potential for local communities to engage in data collection about nearby landfills can lay the foundation for not only greater understanding of the geospatial presence of landfills in the US, but also create a more open and transparent waste management system with the potential to reduce the negative social and environmental impacts of waste collection.

While tremendous opportunities exist for the use and incorporation of user-generated data, there are numerous challenges that must be overcome. Unlike other forms of urban data, however, user-generated data offers not only a unique opportunity to collect novel information and crowdsource the analysis of large datasets, but also opportunities for innovative offline collaboration and active participation in government processes. In order to achieve these outcomes, however, significant effort is required to develop the necessary data collection platform with careful consideration of user interaction and to integrate, process and validate user-generated responses.

CHAPTER 5

Evolving Technologies for Data Acquisition: The Development and Evaluation of Commodity Hardware for Monitoring PM2.5

The proliferation of low-cost sensor technologies has motivated hyper-local sensing initiatives capable of providing new streams of real-time data about numerous environmental and urban phenomena. Given the reduced cost and improved computing power of these devices, dense sensor networks are being used in a variety of applications ranging from transportation monitoring to building-level energy monitoring. While these initiatives vary significantly in scale, objective and methodology, each fundamentally relies on acquiring quality data from the sensor platform. In order to do so, careful calibration and evaluation of the hardware and software technologies is required.

In this chapter, we present the design of a low-cost air quality monitoring platform based on the Shinyei PPD42 (PPD42) aerosol monitor and examine the suitability of the sensor for deployment in a dense spatial network configuration. We assess the sensor’s performance during a field calibration campaign from February 7th to March 25th 2017, with a reference instrument in New York City and present a novel calibration approach using a machine learning method that incorporates publicly available meteorological data in order to improve the sensor’s performance.

This work is a part of a long-term study, the Quantified Community, aimed to understand neighborhood-scale interactions between the environment and man-made infrastructure and their effects on individuals and communities. To understand this complex interaction, a dense sensor network is being devel-

oped to collect granular real-time spatiotemporal environmental data. The air quality monitoring platform described in this work is currently being deployed throughout neighborhoods in New York City and just one aspect of a multi-sensor platform being developed.

5.1 Methods

5.1.1 Node Design

The sensor platform was developed using commodity hardware and designed to capture environmental parameters including fine particulate mater, ambient noise level, air temperature, relative humidity and luminosity. To achieve a high density monitoring network, the selection of sensors and platform hardware required careful consideration in order to find a balance between performance, reliability, accuracy, cost and scalability. Our sensor platform is designed to be deployed in a variety of urban environments, including dense, high-rise neighborhoods with comprehensive digital infrastructure to low density, economically disadvantaged communities with incomplete access to power and wireless network connectivity.

The PPD42 was selected to measure PM2.5 because of its low cost, ease of use, and performance capability demonstrated in previous work [41, 36, 51, 6, 48, 92]. The PPD42 uses a light scattering technique to estimate particle concentration and is capable of measuring particles greater than 1 μ m in diameter. Particles pass through a lighting chamber where the combination of a light emitter and photodiode detector measure the amount of light scattered by particles passing through the chamber. A 0.25W thermal resistor, located at the bottom of the sensing chamber, increases the air temperature inside the chamber relative to the surrounding outside air temperature to create an updraft that draws particles into and through the chamber.

The PPD42 generates two output signals in the form of digital pulses that are referred to by the manufacturer as Low Pulse Occupancy (LPO) and are

proportional to particle count concentration. In order to distinguish particle size, output P1 is used to measure particles greater than $1\mu\text{m}$ and output P2 is used to measure particles greater than $2.5\mu\text{m}$. Particles with a diameter between $1\mu\text{m}$ and $2.5\mu\text{m}$ are determined by subtracting P2 from P1. The PPD42 outputs are connected to the interrupt pins (INT0 and INT1) of an Atmega microcontroller in order to accurately capture pulses that range from 10-90 milliseconds in length. The raw sensor output is converted into LPO readings and sent to a Raspberry Pi microcontroller via USB every 10 seconds to be stored locally. Though the Raspberry Pi is capable of transmitting the data to a central server for real-time processing, there was no available Wi-Fi connectivity in the study area.

A factory calibrated Bosch SHT31 sensor was used to measure air temperature and relative humidity with an accuracy of $\pm 0.3^\circ\text{C}$ and $\pm 2\%$ relative humidity. The electronics were contained in a 6"x4"x2" gray ABS plastic enclosure with a 5VDC fan attached to the bottom in order to draw air into the enclosure through a 1 1/2" filtered vent. Based on the manufacturer specifications, we estimate complete air exchange inside the enclosure occurs approximately three times per second.

The PPD42 sensor used in this study cost approximately \$15 USD. Additional sensors, the microcontroller platform and enclosure materials added an additional \$80 USD resulting in an overall cost of approximately \$100 USD, which is several orders of magnitude less than reference instruments operated by state and federal agencies.

5.1.2 Reference Instrument

The reference instrument for this study was a Thermo Scientific tapered element oscillating microbalance (TEOM) 1400 that provides continuous PM2.5 mass measurements at hourly intervals. TEOM instruments employ a size selective inlet that accumulates particles on a sampling filter located atop of an oscillating element whose resonant frequency changes proportionally to particle mass [58,

5]. The device is operated by the New York State Department of Environmental Conservation (DEC) and costs approximately \$30,000. Data from the reference instrument was obtained directly from the DEC¹. It was observed that the data contained 32 observations with negative values due to the processing procedure performed by the DEC and were subsequently removed from the analysis.

5.1.3 Study Location

The study site was located at an elementary school (PS 104) rooftop on Division Street in Lower Manhattan. The location is a dense urban area with varying infrastructure comprised of approximately 11% commercial buildings, 10% residential buildings, 22% mixed residential and commercial and 2% industrial buildings within 1000m. Of important note, the site is located less than 50 meters from the Manhattan Bridge with an average of 115,000 vehicles crossing every day [70]. The study area also contains approximately 56 buildings that use oil boiler systems, which are known to be significant sources of particulate matter in New York City [23].

The individual nodes were fixed on a custom mounting platform at a height of approximately 1.5m above the rooftop (approximately 12m from ground level) and 3m from the rooftop edge. The design of the mounting platform positioned two devices facing east towards the Manhattan Bridge and one device facing west away from the bridge. The devices were located approximately 5m from the intake of the reference instrument due to logistical reasons.

5.1.4 Performance Evaluation

An initial evaluation of the PPD42 was conducted to assess the accuracy and precision of the three individual deployed devices. Raw LPO readings were aggregated to an hourly average in order to match data from the reference monitor, and pairwise plots were used to compare individual sensor responses with the reference monitor. To evaluate the linear relationship between individual

¹www.dec.ny.org

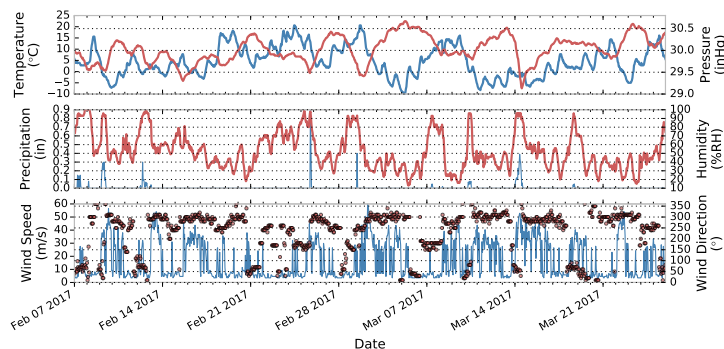


Figure 5.1: Meteorological measurements taken from La Guardia airport over the study period. (a) Temperature (blue) and sea level pressure (red), (b) precipitation (blue) and humidity (red line), and (c) wind speed (blue line) and wind direction (red points).

devices and the reference monitor, an Ordinary Least Squares (OLS) regression was performed on the matched hourly data and R^2 and the RMSE values were used to evaluate the strength and accuracy of the relationship. In this study, measurements from the TEOM monitor are used as the dependent variable and measurements from the PPD42 are the independent variable.

A sensitivity analysis was performed using multiple meteorological parameters to determine their potential influence on sensor measurements. The coefficient of determination was used to evaluate the strength of the relationship between meteorological parameters (independent variables) and the PPD42 and TEOM measurements (dependent variables). Temperature and humidity measurements were taken directly from individual sensor platforms using the SHT31 sensor located inside the enclosure directly adjacent to the PPD42. Other meteorological parameters were also assessed including barometric pressure, wind speed, dew point and precipitation. These measurements were obtained from a nearby weather station located at La Guardia airport. Figure 5.1 shows the meteorological conditions during the study period.

In order to determine the device's sensitivity in low concentration environments, the lower limit of detection (LOD) was calculated as:

$$LOD = 3\sigma_{blk} * \beta_1$$

where σ_{blk} is the standard deviation of the PPD42 measurements obtained when TEOM measurements were below $5.0\mu\text{g}/\text{m}^3$, $3.0\mu\text{g}/\text{m}^3$ and $1.0\mu\text{g}/\text{m}^3$, and β_1 is the slope of the line obtained from the OLS regression analysis. We include multiple calculations of the LOD in order to provide statistically significant results given the small number of samples from the TEOM below $1.0\mu\text{g}/\text{m}^3$ (14 samples). This approach was established by [49] and also used in similar studies [6, 92, 51].

5.1.5 Calibration Approaches

The objective of calibrating the PPD42 is to determine a transfer function that will convert the raw PPD42 sensor readings to match the output of the reference instrument as closely as possible. In order to develop a best-fit calibration model, we evaluate three statistical techniques and incorporate meteorological data to inform and improve our calibration model. All three models were based on measurements from the individual sensor platforms, as well as meteorological data that included air temperature, relative humidity, barometric pressure, dew point and precipitation. As noted in previous work, the PPD42's response is non-linear across the entire range of the device and therefore a quadratic term was also included into the model [36, 6, 92]. A final parameter was added to account for the time of day based on an analysis of diurnal readings from the PPD42 devices, which showed a 1.5 standard deviation difference between the reference instrument during the afternoon hours from 10:00-15:00. This difference is likely caused by solar radiation affecting the sensor's optics and the inclusion of a time parameter is intended to capture this phenomenon. R^2 and RMSE were used to compare calibration accuracy.

The first calibration method used a standard multiple linear regression model in the form of:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

where y is the reference instrument values, β_0 is the intercept, $x_1 \dots x_p$ are the predictors including the PPD42 measurements and ϵ is the error term. The model was specified using best-subset selection, which iteratively finds the combination of features that result in the greatest reduction in the residual sum of squares for each subset of size k where $k = p - 1 \dots p$. The single best model from $M_0 \dots M_k$ was chosen based on Bayesian Information Criterion scores. To detect and account for multicollinearity between variables, the variance inflation factor (VIF) was calculated for all features, and the feature with the highest score was removed. This process was performed recursively until all features' VIF scores were below the threshold of five. The final model included only statistically significant features.

The second calibration technique used a regularization method to address some of the problems with least squares regression. Regularization adds a penalty term (λ) to large model coefficients in order to reduce multicollinearity between features. The ridge regression model used here, applies an ℓ_2 penalty to the sum of the squared coefficients. Ridge coefficients ($\hat{\beta}^R$) are values that minimize:

$$\sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p \beta_j^2$$

where λ controls the amount of penalization. The λ parameter was determined through a five-fold cross validation and set to 0.4. In order to evaluate the significance of individual features, we rank each feature based on the absolute value of the coefficient (β_j). The larger the coefficient, the larger the impact on the model and hence the greater significance of the feature.

The final calibration approach used a GBRT model. GBRT is a decision tree-based regression model that implements boosting to improve model performance. Boosting is a statistical technique that sequentially builds many 'weak'

models (learners) that are combined into a final consensus model [79]. A ‘weak’ learner is one whose performance is only slightly better than random guessing. The final model is built in an additive forward stagewise manner where at each step a new learner is added that minimizes the negative gradient by least squares. The residuals of the current model are then used as the input for the next tree allowing the model to ‘learn’ from the errors of the previous models [33].

Parameter tuning is an important element to optimize the model’s performance. Tree-specific parameters include the depth of each tree, the minimum number of samples to form a terminal node (leaf), and the maximum number of features included in each tree. Boosting parameters include the number of trees used in the model and the contribution of each tree to the final model (learning rate). Tree depth, the number of trees, and the maximum number of features in each tree control the degree of interaction between features. Since trees are grown sequentially, a large number of shallow trees are preferred in order to fully explore the feature space, at the expense of computation time. The learning rate and the minimum number of samples per leaf are used to control overfitting. A low learning rate is generally preferred, but will require a larger number of trees to maintain performance.

To build the ridge and GBRT models, data were first randomly split into train (80%) and test (20%) sets. The training set was used to evaluate model parameters through an exhaustive grid search with 5-fold cross-validation and the final model was evaluated on the test set. All three models were implemented using the scikit-learn package for Python [73].

5.2 Results and Discussion

All three platform nodes collected data continuously throughout the 47-day study period with the exception of four days in which all three devices experienced a power outage. Figure 5.2 shows pairwise plots from the co-located

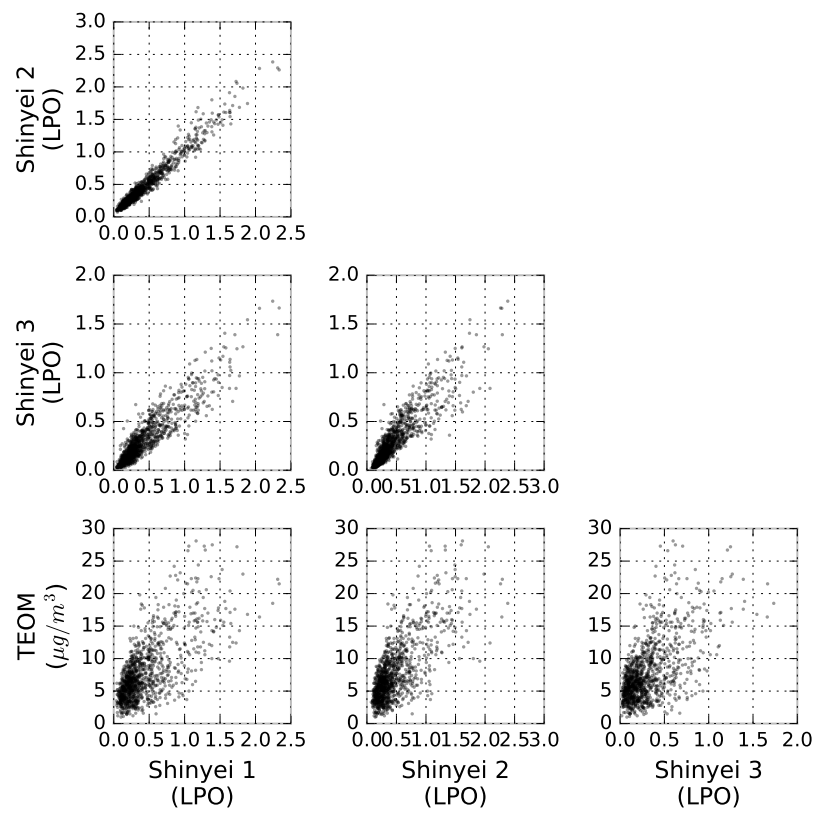


Figure 5.2: Pairwise plots between three PPD42 devices and a reference TEOM based on hourly data collected from February 7th 2017 to March 25th 2017.

PPD42 devices. A total of 1128 hourly observations were recorded from all three devices. Hourly PM2.5 measurements from the TEOM ranged from $1\mu\text{g}/\text{m}^3$ to $28.1\mu\text{g}/\text{m}^3$ with an average of $7.8\mu\text{g}/\text{m}^3$.

Figure 5.3 shows a scatter plot of the linear fit model between the TEOM and PPD42 devices. Based on the calculated R^2 values, individual PPD42 devices demonstrate a moderate level of agreement compared to the TEOM with R^2 values of 0.48 and 0.53 for two devices and the third device slightly lower at 0.37. These results are similar to previous work by [41] who conducted an eight-day field calibration campaign at a regulatory site in Oakland, California and found that a linear correlation was sufficient to explain 55-60% of the variance (RMSE=3.4-3.6) in the federal equivalent method instrument at a one hour interval and 72% at a 24 hour interval. [51] also found moderate correlation ($R^2=0.59-0.8$) between the PPD42 and a commercial grade optical device (TSI DustTrak II Model 8532) during ambient wind tunnel tests, and [36] found similar correlations ($R^2=0.53$) with 24h gravimetric measurements during a four-day calibration campaign in Xi'an, China. [36], however, also observed significantly higher hourly correlations ($R^2=0.87-0.88$) with the DustTrak instrument and suggest the higher correlation is likely due to the increased levels of PM2.5 concentrations observed in Xi'an (range: $77-889\mu\text{g}/\text{m}^3$) compared to [41] (range: $0.3-30\mu\text{g}/\text{m}^3$) since the PPD42's measurement errors increase at lower concentration levels.

Individual PPD42 devices show high correlation with R^2 values of 0.93-0.96 and a linear response across the concentration range. This high correlation between PPD42 devices has been largely consistent across studies by [41], [36] and [51], who all report high inter-device correlations ($R^2 > 0.9$) with the exception of one experiment by [51] reporting a correlation of $R^2=0.72$.

5.2.1 Ambient Conditions

The average temperature during the study period was 4.5°C (range: $-10.0-20.6^\circ\text{C}$) with an average humidity of 52% (range: 0-100%). Rapid fluctuations

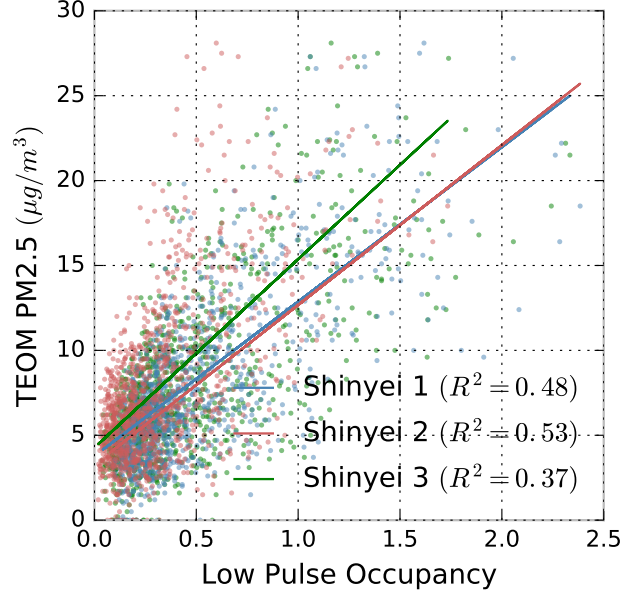


Figure 5.3: Linear model fit for hourly data collected from three PPD42 sensors and a TEOM reference monitor between February 7th 2017 and March 25th, 2017.

in meteorological conditions were observed throughout the study period. For example, the average temperature during the week of February 9th-17th was 0.8°C (range: -7.2-8.2°C) and increased significantly to an average temperature of 10.7°C (range: 1.7-20.6°C) the following week. Other extreme weather conditions were also observed including 20 days with high winds (>30m/s), three separate snow days with a total accumulation of five inches and two days with freezing rain. The observed ranges in temperature, humidity, and precipitation are significantly greater than those of previous field calibration studies.

Table 5.1 shows the sensitivity test results. Dew point temperature measurements show the highest correlation between both the PPD42 and the TEOM ($R^2 = 0.38$ - 0.41 and $R^2 = 0.18$) compared to other meteorological parameters. Temperature and relative humidity are both weakly correlated ($R^2 = 0.24$ - 0.25 and $R^2 = 0.13$ - 0.19) with the PPD42 measurements, and show only minor influence on the TEOM ($R^2 = 0.15$). Previous work by [41] evaluated the effects of temperature, relative humidity and light levels on PPD42 measurements and

Table 5.1: Results of a sensitivity test to evaluate the relationship between meteorological conditions and the Shinyei PPD42 sensor response.

Parameter	R^2			
	Shinyei 1	Shinyei 2	Shinyei 3	TEOM
Temperature	0.25	0.24	0.30	0.15
Humidity	0.19	0.18	0.13	0.03
Dew Point	0.41	0.38	0.38	0.18
Sea Level Pressure	0.01	0.01	0.00	0.02
Wind Speed	0.10	0.11	0.09	0.10
Gust Speed	0.10	0.11	0.09	0.09
Wind Direction	0.12	0.11	0.12	0.02
Precipitation	0.00	0.00	0.00	0.00

found only relative humidity had minor correlation ($R^2 = 0.25-0.28$). While we observe the effect of relative humidity to be slightly lower and the effect of temperature to be significantly higher than findings by [41], it should be noted that the meteorological conditions during the [41] study varied significantly from this study with temperatures ranging from 20 to 30°C and relative humidity ranging between 10-60%. [36] also found that temperature and relative humidity effects were significant, noting the differences in meteorological conditions between their work and findings by [41].

Differences between these studies may be explained by the convective technique used to create air flow through the sensing chamber. Since the convective flow generated by the resistor is proportional to the surrounding air temperature, fluctuations in ambient temperature will have a direct affect on the sensor's ability to draw particles through the sensing chamber. As observed in this study, and noted by [36] and [51], cooler ambient temperatures will more significantly affect the PPD42 measurements than higher ambient temperatures. Furthermore, [51] also compare the PPD42 with a similar optical aerosol monitor, the Plantower PMS3003, and suggest that the improved performance of the PMS3003 may be due to the use of a fan to control air flow through the sensing chamber.

Table 5.2: Results from calculating the lower limit of detection for the PPD42 during a field calibration campaign with a TEOM reference instrument. Units are in $\mu\text{g}/\text{m}^3$

Concentration	Sample Size	Shinyei 1	Shinyei 2	Shinyei 3	TEOM
$< 1\mu\text{g}/\text{m}^3$	14	3.34	2.90	2.30	0.79
$< 3\mu\text{g}/\text{m}^3$	90	3.35	3.30	4.45	2.75
$< 5\mu\text{g}/\text{m}^3$	323	4.82	4.65	5.12	3.37

5.2.2 Limit of Detection

Table 5.2 shows results for the PPD42’s lower limit of detection. The average LOD is $4.83\mu\text{g}/\text{m}^3$ for concentrations below $5.0\mu\text{g}/\text{m}^3$ (323 samples), $3.6\mu\text{g}/\text{m}^3$ for concentrations below $3.0\mu\text{g}/\text{m}^3$ (90 samples) and $2.8\mu\text{g}/\text{m}^3$ for concentrations below $1.0\mu\text{g}/\text{m}^3$ (14 samples). These findings are in the range of laboratory tests performed by [6] ($1.0\mu\text{g}/\text{m}^3$) and [92] ($4.59\mu\text{g}/\text{m}^3$ and $6.44\mu\text{g}/\text{m}^3$).

5.2.3 Calibration Results

Table 5.3 and Figure 5.5 compare OLS, Ridge and GBRT results from the hourly test data and show that the GBRT model significantly outperforms both the OLS and Ridge models with an average R^2 of 0.72. While it is expected that the more complex model will outperform other models, there are two observations that should be highlighted. First, the overall magnitude of improvement by the GBRT model is significant, increasing by approximately 20-30% over the Ridge model. Second, the GBRT model also reduces the range of scores between devices from 0.16 points in the Ridge model to 0.08 points in the GBRT model. This ability to reduce device variability is a significant enhancement for relative calibration and large-scale deployments.

Figure 5.6 compares OLS, Ridge and GBRT calibrated hourly measurements. Overall, the OLS and Ridge models show similar R^2 values and track well against the TEOM monitor. However, results from the OLS and Ridge models periodically under- and overestimate TEOM measurements. Significant under-estimates by the PPD42, for example, are observed on February 11th

Table 5.3: Comparison of results from three calibration techniques.

Parameter	OLS				Ridge				GBRT			
	R^2	RMSE	β_0	Slope	R^2	RMSE	β_0	Slope	R^2	RMSE	β_0	Slope
Shinyei 1	0.452	3.28	3.60	0.59	0.466	3.24	3.35	0.62	0.716	2.36	1.84	0.79
Shinyei 2	0.507	3.11	3.28	0.64	0.521	3.07	2.99	0.67	0.762	2.16	1.47	0.83
Shinyei 3	0.360	3.55	4.74	0.44	0.364	3.54	4.31	0.48	0.678	2.52	2.48	0.72

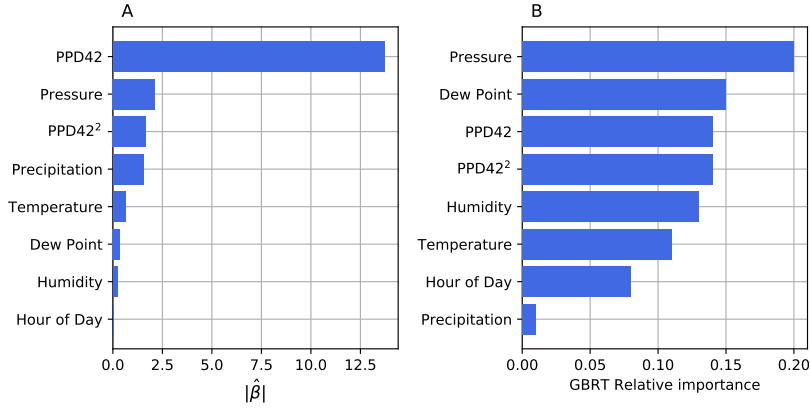


Figure 5.4: Feature importance for the ridge regression model (A) and the gradient boosting regression model (B).

and February 16-19th, in which the TEOM instrument reported higher PM2.5 concentrations during both periods. Over-estimates are often found during the evening hours (e.g. Mar 9-12th) and are likely due to the low PM2.5 concentration levels that fall below the PPD42's lower limit of detection. The GBRT model, however, does not demonstrate the same under- and over-estimates observed in the OLS and Ridge models.

Figure 5.4 compares feature importance between the ridge model and GBRT model. The most significant features in the ridge model are the PPD42 output, sea level pressure, and the squared PPD42 sensor output, while the GBRT model identifies pressure, dew point, the PPD42 output and the squared PPD42 sensor output. These results also show that the Ridge model places greater weight on only a few parameters, while relative feature importance is distributed across features in the GBRT model. This is expected given that the GBRT model is a more robust model capable of learning complex relationships across a large set of input parameters. In this case, the model is able to better establish

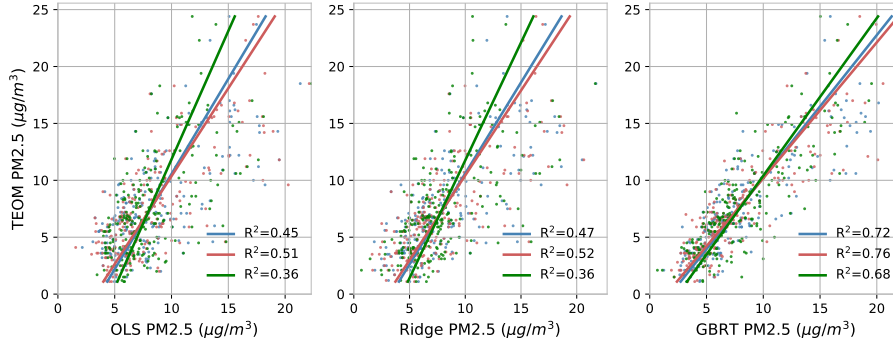


Figure 5.5: Scatter plots of three Shinyei PPD42 sensors calibrated with three different techniques. Sensors are calibrated through a multi-linear regression, ridge regression and gradient boosting regression tree model.

the relationship between sensor measurements and meteorological conditions to improve the calibration.

5.2.4 Main findings

The aim of this study is to examine the viability of a low-cost air quality platform based on the PPD42 aerosol monitor to measure PM2.5 in a dense urban environment. Based on an extensive field calibration campaign, we find the PPD42 performs reasonably well throughout a variety of environmental conditions and can be a suitable device for measuring PM2.5, especially considering the difference in cost from other commercially available instruments. The high correlation between PPD42 devices is particularly significant for high-density sensor networks that rely on relative measurements to inform the spatial distribution and variability of PM2.5 across a study area. Furthermore, while measurement errors increase at lower PM2.5 concentrations ($< 5 \mu\text{g}/\text{m}^3$), the limit of detection falls below the range of ambient concentration levels expected in many urban environments. For example, New York City’s average annual PM2.5 concentration level is $11.55 \mu\text{g}/\text{m}^3$ with a range of 5.17 - $26.48 \mu\text{g}/\text{m}^3$ [64].

An important consideration in evaluating acceptable detection limits is the specific application and use of the recorded particulate matter observations. For example, large absolute measurement errors from low-cost devices may still

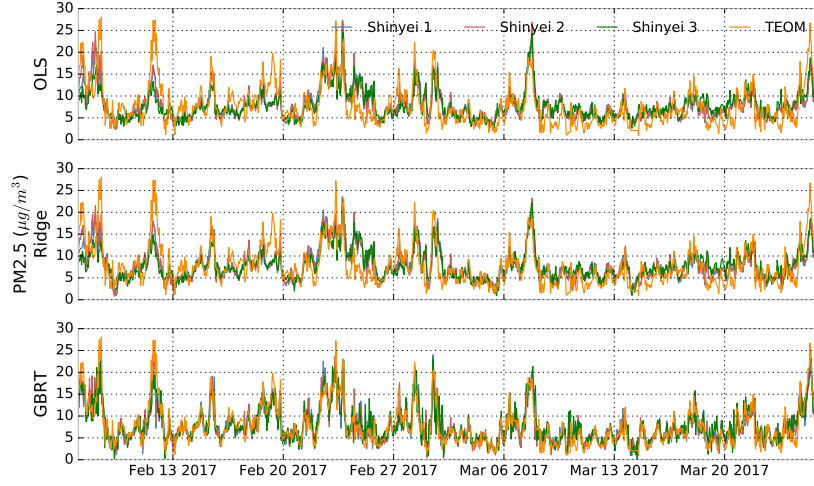


Figure 5.6: Comparison of calibration results with a reference instrument using different calibration techniques including multiple linear regression, ridge regression and gradient boosting regression tree models. Hourly PM2.5 measurements were obtained from three Shinyei PPD42 sensors co-located with a TEOM reference instrument from February 7th through March 25th, 2017.

be acceptable when comparing ambient PM2.5 levels between communities. In this scenario, the relative performance of these sensing devices can provide the necessary information for comparison. Similarly, a dense sensor network comprised of low-cost devices can be used to determine localized hot spots and PM2.5 sources by comparing relative readings across individual sensors within the network. Not only can low-cost devices provide the necessary information for the detection of these point sources, they are also well suited for large-scale deployments given their low cost. In this configuration, low-cost sensing devices can inform local communities and city agencies about baseline patterns of PM2.5, as well as quickly identify anomalous events. Another important aspect when using low-cost devices is their ability to engage local residents by providing hands-on experience measuring air quality but also the ability to quickly provide information to residents. This feedback loop is important for increasing community awareness and providing the necessary actionable information. Finally, the temporal resolution offered by many low-cost devices, including the PPD42, can be useful in measuring transient emission sources that may significantly exceed

ambient concentration levels over short time periods.

Through comparing various calibration techniques, this study found that a GBRT model that uses publicly available meteorological data can significantly improve the performance of a low-cost aerosol monitor. While this calibration process does not necessarily establish an equivalence between the devices, it does provide a method for converting raw sensor readings into standard units ($\mu\text{g}/\text{m}^3$) and improve the sensor’s performance by identifying meteorological conditions that cause measurement error and adjusting the sensor’s response accordingly. Furthermore, the implementation of a ML model to calibrate low-cost instruments can be a step towards a universal calibration curve and standardize sensor deployments. A properly trained ML model could be publicly distributed and implemented in similar hardware deployments by citizen science communities and nonspecialists, which could reduce the need to calibrate devices individually, improve long-term device stability, and standardize data generation and collection methods.

5.2.5 Limitations

A significant limitation when using the PPD42 is the inability to explain measurement errors and variability between the PPD42 devices. This is largely a result of the optical sensing technique employed. Unlike other sampling techniques, the light scattering approach used by many low-cost aerosol monitors is unable to evaluate the physical properties of particles such as composition, type, mass, or optical characteristics. For example, organic particles tend to absorb moisture from the surrounding environment making them more susceptible to changes in humidity. Similarly, different particle types have different optical properties that can vary depending on the wavelength of light used in the sensor.

This work is also limited by the use and comparison of three sensing units, which limits a full evaluation of inter-device variation. Though our analysis is consistent with previous work showing high correlation ($R^2=0.93-0.96$) between

PPD42 devices, a more robust statistical analysis that includes greater than 10 devices has yet to be performed. Similarly, while the calibration campaign does provide sufficient data to assess the sensor’s performance in concentration ranges typical for New York City, these ranges may vary significantly in other urban areas around the world. To ensure accurate calibration, especially when using ML techniques, the devices should be exposed to the entire range of concentrations expected during deployment in order to include the training data necessary for the model to establish the proper input-response relationship. Furthermore, the study duration also limits an evaluation of long-term stability ($>1\text{yr}$) and time-in-use effects such as the gradual accumulation of particles inside the sensing chamber, which may effect the sensor’s optics.

There are also several important limitations to implementing machine learning algorithms for sensor calibration. One significant challenge is the potential to overfit the model to either the specific environment in which the calibration took place, or to the sample data used for the calibration. The latter is a general concern whenever using machine learning models and can be addressed with various techniques such as cross validation, as implemented in this analysis. Overfitting the calibration environment, however, can occur by incorporating parameters into the calibration model that are either specific to the calibration location, or do not include the full range of conditions that the sensor will be exposed to during deployment. It is essential that individual parameters contain sufficient variance to properly capture potential deployment conditions, while excluding any spatial parameters that could potentially affect the input stimulus (i.e PM2.5). During this study, for example, wind direction was observed to explain 10% of the variance of the TEOM monitor and the inclusion of this parameter in the GBRT model improved results on average by 5%. However, the effect of wind direction on PM2.5 in this specific location may result from variations in the built environment that potentially include PM sources (e.g buildings with specific boiler types), which will likely differ from deployment locations. Including wind direction would therefore train the calibration model

based on the specific conditions of the study location instead of identifying the interaction of non-site specific variables that affect the PPD42. Similarly, the inclusion of a time-of-day parameter could lead to erroneous calibration errors since diurnal PM2.5 trends may be affected by local emission sources that vary per location.

Furthermore, while a machine learning model can increase overall performance, it is unable to explain measurement error nor provide information about particle properties. Feature importance is one method to understand how the model is using features to make predictions and adjust the sensor response, but it does not necessarily describe the affect of certain meteorological parameters, or combinations of parameters, on the sensor's response.

5.3 Discussion of Sensor Data

Broadly, this work provides a detailed look at data obtained from a low-cost sensor technology and describes how the data can be manipulated, transformed and integrated in order to generate information about air quality. This study in particular offers specific insight into the abilities and challenges of sensor data, but most importantly, targets the fundamental question about the potential for these devices to generate quality data. Fundamentally, this form of data provides the largest potential for advancing urban informatics studies, but specific applications and uses of these technologies still need careful consideration and evaluation in order to properly develop and leverage the potential provided by these tools. As presented in this work, a detailed understanding of a sensor's technical operation and data output is required, as well as a thorough evaluation and validation of the sensor data, in order to obtain meaningful information.

Furthermore, both the sensor data and calibration data can suffer from data gaps and errors resulting from technical malfunctions, environmental influences and data corruption. In this example, not only did the data used for calibration occasionally experienced failure, the sensor platform itself also experienced

outages during the study period. Identifying these potential failure points are critical for ensuring data quality and preventing the proliferation of networked technologies incapable of providing useful information.

Another important aspect of this work, as seen in the previous two chapters, is the importance of combining various data sources to maximize the potential of urban data. In particular, the calibration of the sensor platform was based on an administrative data set and the fundamental efforts to improve the calibration model were based on the incorporation of weather data obtained through available open data sources. Simply put, administrative data were essential for enabling the research and advancing the findings.

While there are important limitations to sensor data, it is also important to weigh the potential opportunities made available by these devices. In particular, it is important to note the stark differences in the sensor technologies used in this study. The sensor platforms developed for this study are estimated to cost several hundred dollars each, while the reference instrument costs tens of thousands of dollars and requires constant maintenance and supervision. This significant cost difference allows for large-scale sensor deployments, which can provide useful information despite the potential reduction in data quality.

Finally, an important characteristic of sensor data often overlooked is the potential threat to public privacy. The proliferation of technologies capable of collecting detailed information about individual activity presents new challenges to prevent these technologies from being used to obtain or identify aspects of individuals' lives that could potentially violate individual privacy. As with other data types, the integration of multiple data sources, including sensor data, can result in the identification of individual behavior patterns that can be deemed intrusive and harmful to individuals. Avoiding these scenarios from a sensor perspective requires not only careful consideration of the sensor technologies being implemented, the locations and frequency of data collection, and the secure transfer and storage of these data, but also secure data sharing and access practices, and thoughtful approaches towards integrating, processing and

analyzing sensor data throughout the lifecycle of the research.

5.4 Chapter Summary

This study evaluates the usefulness of sensor data through a demonstration of the suitability of a low-cost aerosol monitor to measure intra-urban PM2.5 concentrations. Over a 47 day study period, three PPD42 sensors, integrated with a Raspberry Pi microcontroller and Bosch SHT31 temperature and relative humidity, were deployed on the roof of an approximately 12m high building proximate to a TEOM instrument installed and operated by the New York State DEC. The devices were exposed to wide variations in ambient temperature, relative humidity, barometric pressure, and precipitation in an environment characterized by a diversity of urban land use types. Potential point sources of pollution included 56 surrounding buildings using oil boilers for heating and the vehicular traffic along the Manhattan Bridge.

We evaluate three machine learning methods to calibrate the deployed sensors, including traditional OLS regression, Ridge regression, and a GBRT decision tree model. Our results indicate that the GBRT model significantly outperforms the OLS and Ridge models. Overall, we find that low-cost aerosol devices can be reliably used for community air quality monitoring defined by a dense spatial network of sensors in heterogeneous urban environments. The GBRT calibration method provides reasonable performance when combined with meteorological data that can be used to convert raw sensor readings to standard units. Importantly, this machine learning approach can also be used to develop a universal calibration curve to standardize readings across field-deployed sensors.

Specifically, this chapter provides three main contributions directly related to air quality sensing in cities using commodity hardware. First, this research contributes a new method for integrating the hardware components into a robust sensing platform. Although the components themselves are not new, the technical approach to integrating these devices provides a new approach to collecting

data about air quality. Furthermore, the detailed description of the hardware components in this chapter offers new insight into the operation of these devices and how data is being generated by the sensors, including a description of how the sensor can be affected by diverse environmental conditions. Previous literature related to hardware device operation often lacks a clear technical description which is essential when assessing a device's capabilities and performance.

The second significant contribution in this chapter is a new understanding of the performance of a low-cost air quality monitor. Specifically, the PM2.5 sensors used in this research are assessed under environmental conditions not previously assessed in other studies. The calibration campaign presented in this chapter is conducted not only over a significant time period which is important for generating the information necessary to provide statistically significant results, but also exposed the sensors to a wide range of conditions which had not previously been considered. By testing the performance in these conditions, this research is able to provide new insight into how the sensors perform and critically, what environmental factors will influence sensor readings the most.

The third contribution presented in this chapter is a new method for calibrating these sensors in order to improve their performance in an urban environment. Previous studies have focused on evaluating the linear relationship between the test device and reference instrument through multiple linear regression techniques. While these approaches are important and fundamental ways to critically assess these instruments, this research presents an alternative method using an external source of information and a robust machine learning technique to formulate a transfer function that best describes the relationship between a range of environmental variables and the sensor output. Through this calibration approach, not only does the overall performance of each sensor improve, but the collective variation between individual devices narrows. While there are drawbacks to this form of calibration, the use of a more advanced statistical technique to calibrate a low-cost devices demonstrates the potential to

significantly improve the performance of these devices, as well as the potential to generate a universal calibration curve that could be distributed to air quality sensing communities in order to standardize sensor performance across multiple deployments and studies.

Finally, this chapter discusses the unique opportunity generated through the use of low-cost technologies to generate high-resolution sensor data. Despite the potential drawbacks in the quality of data generated, the application of these technologies can provide novel and high-resolution information important to understanding the urban environment, capturing dynamic changes in patterns of urban activity and generating real-time streams of information capable of providing insight into complex and challenging urban phenomena. Sensor data, therefore, represents one of the largest opportunities in the field of urban informatics, but also faces significant challenges.

CHAPTER 6

Discussion: The Dynamics of Data in Urban Informatics Research

The central argument in this thesis has developed through three independent urban informatics studies that investigate the effectiveness of specific data sources and tools in order to better understand socio-ecological relationships in cities. Chapter 3 demonstrated the ability of an administrative dataset to successfully predict waste generation for short time scales and therefore improve operational efficiencies, while Chapter 4 highlights user-generated data, which can provide information when administrative datasets are insufficient. Chapter 5 highlighted in situ sensor data by developing and evaluating a low-cost air quality monitoring platform and improving its performance by using ML techniques for calibration. Independently, these studies demonstrate important characteristics of the specific urban data types and provide insight into the complex socio-ecological relationships that exist in cities. When viewed collectively, these studies not only reveal several important aspects of urban informatics research that are often overlooked and considered as peripheral components in quantitative urban studies but also lay the foundation to develop a framework for the quantitative study of socio-ecological relationships in cities - an ecological urban informatics.

This chapter examines the three studies in this thesis as a collective whole and discusses the important insights these studies provide about socio-ecological relationships and urban systems, as well as selected examples of the peripheral components of urban informatics including the importance and necessity of working with external organizations, data quality trade-offs, the complementary nature of big urban data and concerns of data privacy and security. The chapter also includes a brief discussion of the important challenges and limitations to

specific studies presented in this thesis, and concludes with recommendations for urban informatics practitioners and a general discussion of the usefulness of new data sources and tools for data acquisition for studies in urban informatics.

6.1 Towards an Ecological Urban Informatics

The individual studies presented in this thesis have been motivated by the need to understand the socio-ecological relationships in order to better understand and mitigate the negative environmental effects of cities and ensure sustainable future cities. While a complete articulation and exploration of socio-ecological relationships is largely outside of the scope of this thesis, it is important to provide a fundamental description of these relationships in order to bring these individual studies together in a cohesive manner. This fundamental overview will also highlight a vision for developing an ecological urban informatics framework that will enable the future study of complex urban dynamics.

The relationship between human activity and the negative effects of this activity on the surrounding environment is becoming increasingly clear. Waste management is an excellent example of this socio-ecological relationship in which society's consumption and disposal habits result in the generation of trash, which can have negative environmental impacts caused by landfilling, incineration and emissions generated from trash collection. The management of waste and its impact on the environment, however, is influenced and driven by numerous factors that range from political and social to economic and logistical. While on the surface the relationship between waste and the natural environment may seem one dimensional and straightforward, in reality, it exists as a complex system shaped by numerous internal and external factors.

The relationship between waste and society is most evident in Chapter 3, which not only quantitatively describes the process of urban waste generation with unprecedented detail but also provides new insight into the regular nature of this process. Specifically, this chapter uses detailed information about

waste collection to describe a city-wide waste generation process that occurs with great regularity including patterns of seasonal waste generation and the collective process of generating waste that occurs during holidays and special events. Furthermore, long-term trends can be identified and explained by social phenomena. For example, the decreasing trend of paper recycling can be attributed to the increased use of digital tools for media and content delivery, therefore decreasing the use of traditional print media and paper products. Similarly, the effects of policy changes and the implementation of new waste-related programs can be clearly identified as seen by the dramatic decline and subsequent increase in recycling that took place as a result of a city-wide suspension of the recycling program from 2002 through 2004.

Chapter 4 elaborates on the relationship between society and waste and uniquely redefines urban boundaries as a result. Through an online crowdsourcing project that focuses on identifying the locations where waste is deposited, participants are able to visually identify the impact and spatial extent of their individual waste generation habits. Figure 6.1 more concretely defines this process by visualizing and directly linking the sources of waste (i.e. DSNY transfer facilities) with the potential destinations (i.e. landfills). While the routes in Figure 6.1 do not represent specific trajectories, they do provide a visual indication of how trash is exported out of New York City and the collective impact that society has on the surrounding region. By spatially visualizing the potential routes that waste could travel, a new perspective about the boundaries and borders of cities is revealed, as well as the complex urban network that extends well beyond the administrative boundaries that traditionally define these areas.

The third study presented in this thesis shifts from waste management to urban air quality as the central socio-ecological relationship. While these two phenomena share many similar characteristics, for example, both process result from human consumption processes, urban air quality extends the understanding of socio-ecological relationships in many ways. Uniquely, urban air quality is a socio-ecological process with well-known and direct links to human health.

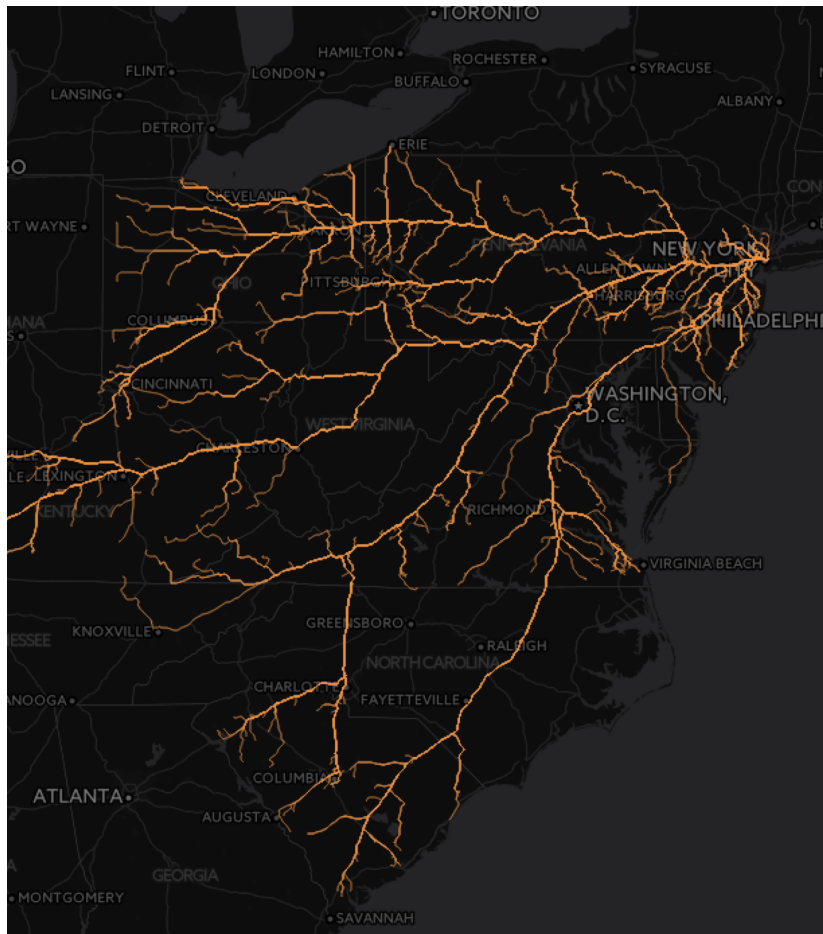


Figure 6.1: A map of New York City's waste disposal network. The orange lines represent possible routes from DSNY transfer stations to landfills located in states known to accept NYC waste. Specific routes were generated by the Google Map API in order to provide specific truck driving routes.

Unlike waste generation, whose impacts are mainly environmentally related and only in certain circumstances has direct links to human health (e.g. leachate that affects drinking water), exposure to poor air quality has a direct negative impact on human health. Although the research presented in this chapter is grounded in the socio-ecological process of air quality, the contributions of this chapter do not directly relate to the phenomena of air quality but rather lay the foundation and infrastructure necessary for future research to measure and provide information about this relationship.

As previously stated, the overall objective of this thesis is not to specifically define, elaborate and advance our understanding of socio-ecological relationships that are mediated and shaped by cities. Instead, this thesis has explored a range of novel tools and techniques and evaluated the potential usefulness of these approaches to study these relationships. In doing so, this research has laid the foundation for developing a new framework for the study of socio-ecological urban systems. Specifically, these three studies have identified a unique research direction that bridges the traditional fields of ecology, urban studies and informatics to create an ecological urban informatics. Ecological urban informatics is, therefore, the science of collecting, processing and analyzing data to study the socionatural interactions and relationships in cities and urban environments. This new framework seeks to leverage new sources of information and new information communication technologies to study, explore and integrate urban systems including infrastructure, communication and transportation systems, natural and environmental systems including air quality, water quality and land use, and population dynamics that include economic, political and social drivers. By integrating across this range of topics, incorporating new technologies and information, and using cities as a unifying lens, this new framework has significant potential to generate novel insights that can not only lead to advanced theoretical insights about urban systems and the dynamics of these systems but also lead to important real-world outcomes for local governments and community organizations, as well as the improved quality of life for urban residents.

6.2 Partnerships and Collaboration

An essential, and often overlooked, aspect of urban research is relationship building. Though some urban research can be performed in isolation, the vast majority of urban informatics research depends highly on working with individuals and organizations to facilitate and implement research initiatives. For example, sensor deployments often face complex technical and socio-political challenges that can range from mitigating impacts of external elements (i.e. weather or human tampering with the device) to legal and privacy issues inherent in collecting information in urban spaces [68]. Many of these challenges can be mitigated through collaboration with local communities, neighborhood organizations and government agencies, who can also provide insight relevant for problem identification, contribute to data collection and analysis efforts, and provide tacit knowledge about the neighborhood useful for understanding and validating results. Indeed, collaboration with external organizations was an essential element for each study presented in this thesis.

Chapter 3 provides a clear example of the benefits of conducting research in partnership with a city agency. Not only did DSNY directly enable the research by providing a rich historical dataset, they also played an important role in contributing detailed information about citywide operations, explaining specific challenges faced by the agency and providing a clear understanding of the source data and how it was generated. The agency’s tacit knowledge of the city’s waste collection system was critical for proper data organization and cleaning processes, as well as provided insight on and confirmation of preliminary analyses. Furthermore, the collaboration with DSNY allows for further experimentation and validation of results through specific DSNY directed efforts, and the creates the possibility for the implementation of data-driven solutions. This work in particular, has laid the foundation for a continued partnership with DSNY who continue implement data-driven methods to improve operations.

The study presented in Chapter 5 was also facilitated through a relation-

ship with a government agency; the New York State DEC. In this case, the agency not only provided the necessary access and support to co-locate the low-cost sensor platforms alongside an agency-operated reference monitor, they also regularly collect and make publicly available the data from the instrument. Though the interaction with the DEC was less formal and more indirect, the collaboration was nevertheless an essential element to conduct the research.

Chapter 4 demonstrates yet another important form of collaboration - public participation. Indeed, the entire project is based on the creation of an online collaborative environment where individuals participate by performing data collection and validation tasks. As highlighted by this study, public engagement is non-trivial: to combine data collection methodologies while capturing public interest and enthusiasm requires not only an understanding of the underlying problems being addressed, but also an understanding of the principals and motivations of the communities involved. Engagement with local communities can have numerous primary and secondary benefits for both urban research and the communities themselves, and the importance of this engagement should not be overlooked. Similarly, engagement at an individual level through online crowdsourcing, participatory data collection efforts and citizen science initiatives, can lead to important educational opportunities, the generation of new sources of information and creative online and offline experiences that ultimately shapes the very fabric of urban life.

6.3 Complementary Roles of Urban Data

The three studies presented in this thesis each focus on the application and usefulness of a specific type or stream of new urban big data. These individual data streams, however, have a significant ability to enhance and complement each other and should not be considered as individual tools to be used in isolation. The integrating and combining of multiple data types can not only provide information superior to an individual stream, but can also provide information

necessary to validate and substantiate a single data source. Furthermore, the characteristics and specific qualities that make a dataset useful for one scenario may not hold true for another scenario. Understanding the strengths of a specific data source and how to use other data sources in a complementary way is essential for the successful use of urban big data.

As previously discussed, administrative data sources have been, and continue to be, fundamental sources of information about cities. In some situations, however, administrative datasets may be insufficient, and alternative sources of information may be required to advance the research objective. For example, the work presented in Chapter 5 is founded on the fact that state and federal air quality monitoring programs in New York City do not provide PM_{2.5} measurements with the resolution necessary to understand spatial variations across short distances. In this case, in situ sensor data may be used to complement and/or supplement existing administrative data. Indeed, the study presented in Chapter 5 uses administrative data to calibrate the low-cost air quality sensors, which will then be used to collect higher resolution PM_{2.5} measurements that can complement and enhance the existing administrative dataset. The study demonstrates that in situ sensor data often requires the use of administrative data in order to generate meaningful data.

Chapter 4 discusses a similar scenario in which the existing administrative data about landfills in the United States were sparse, incomplete and often inaccessible. In this case, the incomplete administrative data was used as a starting point for public participation in which user-generated data not only validated the existing administrative dataset, but also generated and contributed new information back to the administrative dataset. Despite being incomplete, the administrative dataset greatly enhanced the potential for users to identify landfills online.

6.4 Data Quality

Data quality describes the accuracy, coverage and potential sources of error for a particular dataset [85]. To ensure accurate and unbiased analyses, it is essential for researchers to fully understand any potential errors and gaps that may be present in the data. This requires not only a firm grasp of the specific rows and columns that comprise the dataset being used, but also a thorough understanding of how the data were generated, collected and processed. This is especially important when working with data generated by commodity hardware and public participation.

The study in Chapter 5 discusses the considerable measurement and calibration error potentially generated by low-cost sensing devices, often caused by a variety of reasons including the fundamental design and sensing principals. In order to ensure data quality from low-cost sensor sources, a thorough understanding of device operation, the intended operating environment and data collection and processing systems is necessary. For example, minor voltage fluctuation in a sensing platform may have a direct impact on how the sensor operates, which can cause significant variations in the overall output if the voltage is not properly regulated. Furthermore, individual devices will perform differently in different urban environments making it essential to continually evaluate, calibrate and test individual devices [59]. As this study demonstrates, integrating alternative data sources (e.g. weather) and implementing statistical methods can improve the accuracy and performance of low-cost sensing platforms.

User-generated data also faces important data quality challenges. The work in Chapter 4 offers an example of individuals who generated vastly different responses when asked to analyze the same task. These variations may have resulted from data entry errors, technical glitches or the participant's failure to understand the task, and given the large number of potential sources of error, it is essential to implement measures to prevent and identify invalid data. A fundamental approach to validate user-generated data is to distribute the same

task to multiple users and compare the responses. This approach provides a robust statistical way to cross-validate user responses and identify incorrect responses and entry errors. To further extend response validation, participants can be ranked based on their ability to accurately complete tasks, and based on these performance metrics, individual responses can be “weighted”. Interface design and robust user testing also play an important role in successfully engaging participants and preventing data entry errors.

Beyond addressing specific data quality issues related to the various types of urban data, this thesis also raises an important question about the appropriate level of data accuracy and resolution, which is often dependent on the specific research question. For example, data generated by a network of low-cost devices may be used to compare the relative differences across the network in order to understand the spatial variation of PM2.5 across a neighborhood. In this case, a sensing device that demonstrates only minor variation between devices may be preferred over a more costly device that reports measurements with higher absolute accuracy. Alternatively, if the research objective is to understand the affects of PM2.5 exposure on human health, then the precision and accuracy of air quality measurements are critical and the use of lower quality sensing devices may not be preferred. As such, it is essential to critically evaluate the overall research objective in order to determine the types of data and level of accuracy necessary for successful research outcomes.

6.5 Data Privacy

Data privacy is of increasing concern, especially as the number and type of data sources continue to grow, which enhances the possibility of linking seemingly irrelevant datasets that have the potential to reveal detailed and sensitive information about individuals. Indeed, the increased use of social media platforms and proliferation of connected technologies in cities enables new potential to track individual activity both online and offline. While this information can

create new opportunities to enhance services and improve the quality of life of many urban residents, the misuse of these data can lead to significant privacy violations, the dissemination of false information and the targeting of individuals for malicious purposes. Though the work presented in this thesis does not directly employ data that is inherently private or sensitive, it is important to discuss potential scenarios in which the data sources used in this thesis could be used to negatively affect privacy and/or public sentiment. The following is therefore not intended to be a comprehensive review of privacy concerns related to urban data, but a relevant discussion about potential data misuse related to the studies presented in this thesis.

The data used in Chapter 3 describes weekly waste generation for administrative regions throughout New York City. While this information does not directly include private or sensitive information, there are several important aspects about this study that could violate individual privacies if misused. For example, the data used in this study was provided by the New York City Department of Sanitation and in its raw form, contained individual truck collection tonnages along with information about the truck route, as well as other miscellaneous information. From the perspective of understanding and forecasting citywide waste generation, this detailed information was unnecessary and aggregated to a more meaningful level (i.e. weekly waste generation per section). However, if truck routes are properly linked to the specific individuals working those shifts, this detailed information could be used to monitor individual work output, evaluate performance and compare worker productivity. While the detailed monitoring and tracking of individuals may be desirable in certain situations, this information is most likely overly invasive, especially for a public agency.

Similarly, this granular information about waste collection could be used to develop detailed models that estimate building- or household-level waste generation. While this information may be critical for the purposes of developing pay-as-you-throw programs that incentivize waste reduction, they may also lead

to unintended consequences such as invasive monitoring of individual households and families or monitoring waste generated by prominent institutions or individuals.

The study presented in Chapter 4 also presents data privacy concerns in a variety of forms. Given that an individual is generating the source data, there is significant potential and incentive to collect supplementary information about the individual in the process of collecting the primary data, which may be used for the purposes of validating user-generated data and/or assessing the performance of the crowdsourced platform. For example, many crowdsourcing applications will collect individual user activity and track individual performance through the crowdsourcing platform in order to assess a users' ability to provide valid information related to the project. This assessment can be critical to ensure data quality, validate user responses and provide feedback for improving the overall experience of the application. While this information may not necessarily violate users privacy, it is paramount that this information be carefully scrutinized to prevent malicious or unintended use, as well as clearly communicate to the individual that supplemental information is being collected.

The study in Chapter 5 focuses on air quality data generated through a sensing platform. Similar to the previous examples, data about air quality itself does not necessarily pose a significant threat to individual privacy. However, the use of this data and results from analyses can significantly affect individuals and organizations if handled improperly. For example, a study of air quality may provide findings that identify specific neighborhoods or businesses that are either key generators of poor air quality or suffer significantly more than others. In these cases, organizations could suffer from loss of business or receive fines from government entities, while individual neighborhoods could receive unwanted attention or scrutiny resulting in loss of property value and desirability.

6.6 Limitations and Challenges of the Case Studies

Reproducibility is an important challenge most visible in the work discussed in Chapter 3. The collaboration with DSNY and access to an extensive historical dataset, created a unique context that is difficult to reproduce in other cities. Furthermore, New York City itself has many unique characteristics and specific waste management challenges that make it difficult to juxtapose to other cities. Though this work does provide important insight into forecasting waste generation, it is unclear that generalizing the findings and adapting the same methodology will be as successful for other cities around the world. Further work is needed to compare the results with other cities.

As previously discussed, one of the most difficult challenges when working with user-generated data is the amount of participation required to generate the information necessary to provide statistically significant results. In the example of Landfill Hunter discussed in Chapter 4, the project was largely unsuccessful in attracting significant public participation, which ultimately prevented robust validation of user entries and impeded the project's overall objective of identifying all of the landfills in the United States. Despite the low amount of public participation, however, the project was able to collect enough quantitative information to demonstrate the ability of crowdsourcing efforts to generate data, as well as the potential analytical and exploratory uses of the user-generated data. Furthermore, the project unexpectedly generated novel learning opportunities, spontaneous offline collaborations and new creative endeavors, which offered unique insight into the secondary benefits of public participation. As such, Chapter 4 is primarily focused on the peripheral benefits and outcomes of crowdsourcing initiatives, rather than the quantitative validation and analysis of user-generated data about landfills.

The study presented in Chapter 5 is a part of an ongoing research initiative that aims to deploy a dense sensor network across multiple neighborhoods in

New York City. While initial sensor deployments are underway, a complete study of neighborhood air quality based on low-cost in situ sensing was outside the scope of work presented in Chapter 5. Instead, the work in Chapter 5 aims to evaluate the viability of using a low-cost aerosol monitors to measure PM_{2.5} in an urban setting, and to understand the abilities and limitations of such devices to generate meaningful information. Absent from this work are the experiences and lessons learned from an initial pilot study conducted in Red Hook, Brooklyn [55]. This pilot study informed our understanding of real-world deployment scenarios and greatly influenced decisions related to sensor selection, hardware enclosure design and deployment strategies.

Finally, a significant challenge that underpinned all of the work in this thesis is the technical requirements for the acquisition, storage, management and analysis of the necessary data. An extensive effort was required to develop, test and validate the data acquisition systems discussed in Chapter 5 and Chapter 6. Sensor systems, in particular, require not only significant hardware and software development to properly collect, process and transmit data, but also significant effort to create a platform that can withstand harsh outdoor environments without interfering or biasing sensor measurements. Similarly, crowdsourcing platforms not only requires the development of an online data collection and processing platform, but also demands careful consideration of user interaction with the platform in order to create an effective interface design to facilitate participation. In general, the creation of these data acquisition systems often requires bespoke development that includes a significant time investment, as well as a thorough understanding of a variety of technical systems and protocols to ensure accurate transfer, efficient storage and secure access to the information being generated.

6.7 Recommendations for Local Governments and Practitioners

Broadly, this thesis has been shaped by conducting applied research in an urban setting. Although many of the experiences and lessons learned are not specifically articulated throughout this thesis, substantial practical knowledge has been gained which may be useful for city governments and practitioners of urban informatics. As such, this section aims to identify and highlight practical insights that can serve as recommendations for urban informatics practitioners.

One of the most important recommendations derived from this thesis is the importance of collecting detailed and granular data. The study presented in Chapter 3 demonstrates how detailed information can be used to not only study and explore a complex system in great detail, but ultimately generate operational efficiencies through visualization and analysis. Fundamentally, this entire research endeavor would not have been possible without the DSNY's continued effort to collect information about waste generation, regardless of the short-term benefits from doing so. The usefulness of administrative sources of information should not be underestimated, and while a single dataset may not seem relevant or informative, tremendous insight can be gained through the integration of multiple datasets.

A second recommendation that emerges from this work applies to a broad range of urban informatics practitioners. As stated earlier in this chapter, collaboration is essential. Relationship building between diverse stakeholders is a fundamental component for successful urban informatics research. This not only includes developing partnerships and relationships between diverse organizations, individuals and institutions, but also across city agencies and departments. Synergies between different groups can often be identified through bottom-up techniques such as workshops and individual networks, as well as through large top-down initiatives. Establishing these working relationships is an essential first step towards structuring research initiatives and identifying

common research objectives.

Similarly, urban informatics research should focus on people wherever and whenever possible. On the one hand, there is significant potential to incorporate citizen participation in research initiatives for the sole purpose of efficiency. Crowdsourcing data collection or analysis has demonstrated improved efficiencies, especially under certain circumstances. However, consideration of public involvement should not be based solely the decision to increase efficiency. As demonstrated in this thesis, there are important secondary benefits that can result from public participation. And while those benefits may be difficult to quantify and therefore justify the effort necessary to incorporate citizens, citizen-led and citizen-focused initiatives have the most potential to transform cities. Individuals and communities comprise the underlying fabric of urban life and are fundamentally the ones who have the most invested in ensuring the success of their city.

A final recommendation focuses on the willingness to experiment and try new approaches. While complex socio-political factors often inhibit local governments from rapid modernization, it should not prevent or hinder a desire to innovate. New technologies and new approaches should be embraced and explored in order to discover new efficiencies and generate novel solutions. Cities are dynamic and ever-changing, which ultimately makes any attempt to devise a perfect solution for each urban problem futile. Instead, practitioners should be willing to test the existing status quo and creatively push the bounds of what can be accomplished. And though this type of effort will inevitably produce many failed results, if experimentation is done responsibly, these failed attempts can quickly lead to better solutions and a better understanding of urban dynamics.

6.8 General Discussion

The work included in this thesis has discussed and explored a variety of data types, methods for analyzing these data and novel tools for generating data.

As seen through these studies, significant insight can be gained through the proper use of these data and tools, though careful consideration is needed for the successful implementation of these techniques. Broadly, emphasis should be placed on properly defining the research question prior to the application of these tools. By carefully considering the advantages and disadvantages of the various urban informatics discipline, as discussed in this thesis, researchers can ensure best practices and proper use of data, as well as maximize the available resources to transform data into actionable insight.

Furthermore, the nascent field of urban informatics offers tremendous opportunity for continued innovation and exploration. The techniques and analytical methods presented in the thesis represent only a small subset of the possibilities available to researchers and continue to quickly evolve as new technologies generate new opportunities and increased need for new analytical approaches. By combining various data types in new ways and exploring innovative uses of low-cost sensor technologies or developing new applications for public participation, researchers will continue to find new solutions and insight for the complex and difficult urban challenges.

A juxtaposition of the work discussed in this thesis with the work of urban planners during the early 20th century can offer important insight and serve as a useful precaution as the discipline of urban informatics advances. Similar to the urban planning efforts in the early 20th century, modern urban informatics research is based on the advent and use of new tools and techniques for understanding how cities function. While the specific differences between the two fields are significant, broadly both fields aimed to achieve a new understanding of cities through the use of the technologies at their disposal in order to improve the daily life of its residents. For the urban planners, an enthusiasm for new technology and the application of these tools without carefully considering the diverse reactionary forces and complex dynamics within cities, ultimately led to unsuccessful initiatives and misguided efforts. In many ways, the current state of the urban informatics discipline parallels urban planners of the early 20th

century in that advances in tools and technologies without careful consideration of the underlying sources, quality, and technical operation used to generate data and extract information, as well as the underlying human and environmental dynamics in cities, can fundamentally change analytical outcomes and affect important decision-making processes. The rise and success of the field of urban informatics, therefore, will largely depend on researchers' ability to acquire, integrate and analyze data, while working closely with a variety of stakeholders that can provide the context and domain experience necessary to fully understand analytical outcomes. Ultimately, urban informatics research requires a tacit understanding of the new possibilities made available through big data and novel technologies, awareness of the current urban context and historical factors that have motivated urban changes, and a keen understanding of the relationship between society, nature and cities, in order to successfully advance an understanding of urban life and create livable and sustainable cities in the future.

CHAPTER 7

Conclusion

This thesis has discussed and demonstrated the potential applications of new big data streams to study socio-ecological urban systems. Individually, these works have explored the various types of urban data including administrative data, user-generated data and sensor data, and have generated important contributions related to forecasting waste generation, creating learning opportunities through public participation and crowdsourcing, and the design and calibration of a low-cost air quality monitoring platform. These works have not only leveraged existing tools and data sources for quantitative analysis, they have also developed new tools for data acquisition including online and offline data collection systems, as well as implemented modern statistical and machine learning techniques to demonstrate the usefulness and importance of data-driven modeling.

Collectively, these empirical studies explore the breadth and depth of urban data and highlight important challenges that should be carefully considered when conducting quantitative urban research. Importantly, this thesis identifies a complementary relationship between diverse data types, and highlights the benefits of data integration. The incorporation of multiple data types can be useful for improving spatial and temporal resolution and extending the overall coverage of a dataset, as well as providing a variety of methods to verify data quality and validate research results. Furthermore, these works collectively identify the necessity of establishing and fostering collaborative relationships between diverse stakeholders, including individuals, local communities and government agencies, to promote successful research initiatives.

Given the significance of cities in the 21st century, it is essential to further

our quantitative understanding of urban life in order to create efficient, economic and sustainable future cities. The emerging discipline of urban informatics offers a new set of tools and techniques that can facilitate and greatly improve our understanding of urban environments. By advancing our knowledge of these complex and dynamic urban systems, we can begin to develop novel and data-driven solutions to address the numerous challenges facing cities, and ultimately improve the quality of life for all urban residents.

7.1 Future Work

An important area of future work is the continued expansion and refinement of the urban data taxonomy. While developing a comprehensive urban data taxonomy was not the focus of this thesis, significant work towards further elaborating on urban data types could be conducted. A variety of approaches could be used to classify urban data and a complete urban data taxonomy will constantly evolve over time as new sources of data are made possible through new technologies. The pursuit of this taxonomy, however, has the potential to provide important information useful for understanding the opportunities and challenges of the various data types, as well as identifying the complementary roles and compounding effects of integrating multiple data types as demonstrated in this thesis.

Similarly, this thesis could be expanded through the development and examination of other specific case studies. The work presented here covers only a subset of the field of urban informatics and further studies could be developed to explore other analytical approaches and sub-categories of urban data. For example, further work could examine passive user-generated data sources in order to fully evaluate the difference between active and passive user-generated data, and understand how these sources of information could be complementary. In addition, future urban informatics case studies will not only depend on and contribute to defining an urban data taxonomy, but will also evolve

alongside technological advancements. The discipline of urban informatics has developed rapidly as a result of rapidly evolving technologies, which will therefore require investigation into how these tools and techniques can be used to promote successful socio-ecological research initiatives.

Finally, this work could be greatly extended by examining how urban informatics research affects decision-making processes at the city level. The work presented in this thesis provides specific examples of methods and insights that could inform and improve local government operations, but does not describe how this research is translated and applied by government officials in real-world scenarios. Urban informatics research offers new opportunities to understand the dynamics of cities but, in order to affect positive change, insights must be adopted and implemented by city governments that often face a variety of complex political, social and economic challenges. While these challenges often seem banal, overcoming them can be non-trivial. Significant future work could explore the application of urban informatics research by city officials and evaluate the capacity of these initiatives to generate successful real-world outcomes.

Bibliography

- [1] M. Abbasi, M. Abduli, B. Omidvar, and A. Baghvand. Forecasting municipal solid waste generation by hybrid support vector machine and partial least square model. *International Journal of Environmental Research*, 7(1):27–38, 2012.
- [2] M. Alberti. *Advances in urban ecology integrating humans and ecological processes in urban ecosystems*. Number 574.5268 A4. 2008.
- [3] S. Albeverio, D. Andrey, P. Giordano, and A. Vancheri. *The dynamics of complex urban systems: An interdisciplinary approach*. Springer, 2007.
- [4] F. Albrecht, M. Zussner, and C. Perger. Using Student Volunteers to Crowdsource Land Cover Information. *Geospatial Innovation for Society*, (2014):314–317, 2014.
- [5] S. S. Amaral, J. A. de Carvalho, M. A. M. Costa, and C. Pinheiro. An overview of particulate matter measurement instruments. *Atmosphere*, 6(9):1327–1345, 2015.
- [6] E. Austin, I. Novosselov, E. Seto, and M. G. Yost. Laboratory evaluation of the shinyei ppd42ns low-cost particulate matter sensor. *PloS one*, 10(9):e0137789, 2015.
- [7] L. Barbosa, K. Pham, C. Silva, M. R. Vieira, and J. Freire. Structured open urban data: understanding the landscape. *Big data*, 2(3):144–154, 2014.
- [8] A. Barrett and J. Lawlor. *The Economics of solid waste management in Ireland*. Economic and Social Research Institute, 1995.
- [9] M. Batty. Urban modeling. *International Encyclopedia of Human Geography*, 2009.
- [10] M. Batty. *The new science of cities*. Mit Press, 2013.
- [11] P. Beigl, S. Lebersorger, and S. Salhofer. Modelling municipal solid waste generation: a review. *Waste management (New York, N.Y.)*, 28(1):200–14, 2008.
- [12] L. L. Bergeson. ECHO: Enforcement online, up close, and real personal. *Environmental Quality Management*, 12(4):81–84, 2003.
- [13] M. Bloomberg. A greener greater new york. *The City of New York*, 2006.

- [14] C. Buckley and A. Ramzy. Before Shenzhen Landslide, Many Saw Warning Signs as Debris Swelled, Dec 2015.
- [15] C. Campbell. New york city open data: A brief history, Mar 2017.
- [16] A. Caragliu, C. Del Bo, and P. Nijkamp. Smart cities in Europe. *Journal of Urban Technology*, 18(2):65–82, 2011.
- [17] S. Carnes, M. Schweitzer, and E. Peelle. Measuring the success of public participation on environmental restoration and waste management activities in the US Department of Energy. *Technology in Society*, 20(1998):385–406, 1998.
- [18] N. Castell, F. R. Dauge, P. Schneider, M. Vogt, U. Lerner, B. Fishbain, D. Broday, and A. Bartonova. Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates? *Environment international*, 99:293–302, 2017.
- [19] N. B. Chang and Y. T. Lin. An analysis of recycling impacts on solid waste generation by time series intervention modeling. *Resources, Conservation and Recycling*, 19(3):165–186, 1997.
- [20] Z. Cheng, L. Luo, S. Wang, Y. Wang, S. Sharma, H. Shimadera, X. Wang, M. Bressi, R. M. de Miranda, J. Jiang, et al. Status and characteristics of ambient pm 2.5 pollution in global megacities. *Environment international*, 89:212–221, 2016.
- [21] C.-Y. Chong and S. P. Kumar. Sensor networks: evolution, opportunities, and challenges. *Proceedings of the IEEE*, 91(8):1247–1256, 2003.
- [22] M. J. Clarke and J. A. Maantay. Optimizing recycling in all of new york city’s neighborhoods: Using gis to develop the reap index for improved recycling education, awareness, and participation. *Resources, conservation and recycling*, 46(2):128–148, 2006.
- [23] J. E. Clougherty, I. Kheirbek, H. M. Eisl, Z. Ross, G. Pezeshki, J. E. Gorczynski, S. Johnson, S. Markowitz, D. Kass, and T. Matte. Intra-urban spatial variability in wintertime street-level concentrations of multiple combustion-related air pollutants: the new york city community air survey (nyccas). *Journal of Exposure Science and Environmental Epidemiology*, 23(3):232, 2013.
- [24] C. C. Conrad and K. G. Hilchey. A review of citizen science and community-based environmental monitoring: Issues and opportunities. *Environmental Monitoring and Assessment*, 176(1-4):273–291, 2011.

- [25] M. Danielsson. Mapping informal and formal recycling sites and flows: Nicaragua to Boston, 2014.
- [26] E. Daskalopoulos, O. Badr, and S. Probert. Municipal solid waste: a prediction methodology for the generation rate and composition in the european union countries and the united states of america. *Resources, Conservation and Recycling*, 24(2):155 – 166, 1998.
- [27] B. de Blasio. One new york: The plan for a strong and just city. Technical report, The City of New York, 2015.
- [28] G. Denafas, T. Ruzgas, D. Martuzevičius, S. Shmarin, M. Hoffmann, V. Mykhaylenko, S. Ogorodnik, M. Romanov, E. Neguliaeva, A. Chusov, T. Turkadze, I. Bocheidze, and C. Ludwig. Seasonal variation of municipal solid waste generation and composition in four East European cities. *Resources, Conservation and Recycling*, 89:22–30, 2014.
- [29] M. Dittus, G. Quattrone, and L. Capra. Analysing volunteer engagement in humanitarian mapping: Building contributor communities at large scale. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, CSCW '16*, pages 108–118, New York, NY, USA, 2016. ACM.
- [30] H. Eiselt. Locating landfills, Optimization vs. reality. *European Journal of Operational Research*, 179(3):1040–1049, 2007.
- [31] J. Elith, J. R. Leathwick, and T. Hastie. A working guide to boosted regression trees, 2008.
- [32] G. M. Foody, L. See, S. Fritz, M. Van der Velde, C. Perger, C. Schill, and D. S. Boyd. Assessing the accuracy of volunteered geographic information arising from multiple contributors to an internet based collaborative project. *Transactions in GIS*, 17(6):847–860, 2013.
- [33] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.
- [34] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001.
- [35] B. Fung. The Shutdown is now clogging up the data economy. Thanks, Congress!, 2013.
- [36] M. Gao, J. Cao, and E. Seto. A distributed network of low-cost continuous reading sensors to measure spatiotemporal variations of pm_{2.5} in xi’an, china. *Environmental pollution*, 199:56–65, 2015.

- [37] K. Garcia. Local law 77 of 2013 organics collection pilot program. Technical report, Department of Sanitation, 2014.
- [38] L. Giusti. A review of waste management practices and their impact on human health. *Waste management*, 29(8):2227–2239, 2009.
- [39] N. B. Grimm, S. H. Faeth, N. E. Golubiewski, C. L. Redman, J. Wu, X. Bai, and J. M. Briggs. Global change and the ecology of cities. *science*, 319(5864):756–760, 2008.
- [40] I. Heimann, V. Bright, M. McLeod, M. Mead, O. Popoola, G. Stewart, and R. Jones. Source attribution of air pollution by spatial scale separation using high spatial density networks of low cost air quality sensors. *Atmospheric Environment*, 113:10–19, 2015.
- [41] D. M. Holstius, A. Pillarisetti, K. Smith, and E. Seto. Field calibrations of a low-cost aerosol sensor at a regulatory monitoring site in california. *Atmospheric Measurement Techniques*, 7(4):1121–1131, 2014.
- [42] D. Hoornweg, P. Bhada-Tata, and C. Kennedy. Waste production must peak this century. *Nature*, 502(7473):615—617, 2013.
- [43] I. Iacovides, C. Jennett, C. Cornish-Trestrail, and A. L. Cox. Do games attract or sustain engagement in citizen science?: a study of volunteer motivations. *CHI EA '13: CHI '13 Extended Abstracts on Human Factors in Computing Systems*, page 1101, 2013.
- [44] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*, volume 6. Springer, 2013.
- [45] C. Jennett and A. L. Cox. Eight Guidelines for Designing Virtual Citizen Science Projects, 2014.
- [46] M. Jerrett, A. Arain, P. Kanaroglou, B. Beckerman, D. Potoglou, T. Sahuvaroglu, J. Morrison, and C. Giovis. A review and evaluation of intraurban air pollution exposure models. *Journal of Exposure Science and Environmental Epidemiology*, 15(2):185–204, 2005.
- [47] N. Johnson. Data and the government shutdown, 2013.
- [48] M. Jovašević-Stojanović, A. Bartonova, D. Topalović, I. Lazović, B. Pokrić, and Z. Ristovski. On the use of small and cheaper sensors and devices for indicative citizen-based monitoring of respirable particulate matter. *Environmental Pollution*, 206:696–704, 2015.

- [49] H. Kaiser and H. Specker. Bewertung und vergleich von analysenverfahren. *Fresenius' Journal of Analytical Chemistry*, 149(1):46–66, 1956.
- [50] C. Kellerman and K. Gibbs. 12 things new yorkers should know about their garbage. Technical report, Citizen Budget Commision, 2014.
- [51] K. Kelly, J. Whitaker, A. Petty, C. Widmer, A. Dybwad, D. Sleeth, R. Martin, and A. Butterfield. Ambient and laboratory evaluation of a low-cost particulate matter sensor. *Environmental Pollution*, 221:491–500, 2017.
- [52] R. Kitchin. The real-time city? big data and smart urbanism. *GeoJournal*, 79(1):1–14, 2014.
- [53] F. Koelsch, K. Fricke, C. Mahler, and E. Damanhuri. Stability of landfills-the bandung dumpsite disaster. In *Proceedings Sardinia*, 2005.
- [54] C. E. Kontokosta, B. Hong, N. E. Johnson, and D. Starobin. Using machine learning and small area estimation to predict building-level municipal solid waste generation in cities. *Computers, Environment and Urban Systems*, 70:151–162, 2018.
- [55] C. E. Kontokosta, N. Johnson, and A. Schloss. The quantified community at red hook: Urban sensing and citizen science in low-income neighborhoods. *arXiv preprint arXiv:1609.08780*, 2016.
- [56] P. Korhonen and J. Kaila. Waste container weighing data processing to create reliable information of household waste generation. *Waste Management*, 39:15–25, 2015.
- [57] M. Kubásek and J. Hrebíček. Crowdsourc Approach for Mapping of Illegal Dumps in the Czech Republic. *International Journal of Spatial Data Infrastructures Research*, 8:144–157, 2013.
- [58] P. Kulkarni, P. A. Baron, and K. Willeke. *Aerosol measurement: principles, techniques, and applications*. John Wiley & Sons, 2011.
- [59] P. Kumar, L. Morawska, C. Martani, G. Biskos, M. Neophytou, S. Di Sabatino, M. Bell, L. Norford, and R. Britter. The rise of low-cost sensing for managing air pollution in cities. *Environment international*, 75:199–205, 2015.
- [60] J. Lane, V. Stodden, S. Bender, and H. Nissenbaum. *Privacy, big data, and the public good: Frameworks for engagement*. Cambridge University Press, 2014.

- [61] D. Linders. Towards open development: Leveraging open data to improve the planning and coordination of international aid. *Government Information Quarterly*, 30(4):426–434, 2013.
- [62] Y. Ma, M. Richards, M. Ghanem, Y. Guo, and J. Hassard. Air pollution monitoring and mining based on sensor grid in london. *Sensors*, 8(6):3601–3623, 2008.
- [63] A. Manikonda, N. Zíková, P. K. Hopke, and A. R. Ferro. Laboratory assessment of low-cost pm monitors. *Journal of Aerosol Science*, 102:29–40, 2016.
- [64] T. D. Matte, Z. Ross, I. Kheirbek, H. Eisl, S. Johnson, J. E. Gorczynski, D. Kass, S. Markowitz, G. Pezeshki, and J. E. Clougherty. Monitoring intraurban spatial patterns of multiple combustion air pollutants in new york city: design and implementation. *Journal of Exposure Science and Environmental Epidemiology*, 23(3):223–231, 2013.
- [65] P. R. Mazon. Defiende el territorio desde el aire, 2014.
- [66] M. Mead, O. Popoola, G. Stewart, P. Landshoff, M. Calleja, M. Hayes, J. Baldovi, M. McLeod, T. Hodgson, J. Dicks, et al. The use of electrochemical sensors for monitoring urban air quality in low-cost, high-density networks. *Atmospheric Environment*, 70:186–203, 2013.
- [67] S. Moltchanov, I. Levy, Y. Etzion, U. Lerner, D. M. Broday, and B. Fishbain. On the feasibility of measuring urban air pollution by wireless distributed sensor networks. *Science of The Total Environment*, 502:537–547, 2015.
- [68] R. N. Murty, G. Mainland, I. Rose, A. R. Chowdhury, A. Gosain, J. Bers, and M. Welsh. Citysense: An urban-scale wireless sensor network and testbed. In *Technologies for Homeland Security, 2008 IEEE Conference on*, pages 583–588. IEEE, 2008.
- [69] J. Navarro-Esbri, E. Diamadopoulos, and D. Ginestar. Time series analysis and forecasting techniques for municipal solid waste management. *Resources, Conservation and Recycling*, 35(3):201–214, 2002.
- [70] New York State Department of Transportation. Traffic data viewer, 2017.
- [71] S. Opp. Bureaucratic Discretion and Political Control of the Resource Conservation Recovery Act (RCRA). *International Journal of Public Administration*, 34(12):753–763, 2011.

- [72] I. Oribe-Garcia, O. Kamara-Esteban, C. Martin, A. M. Macarulla-Arenaza, and A. Alonso-Vicario. Identification of influencing municipal characteristics regarding household waste generation and their forecasting ability in Biscay. *Waste Management*, 39:26–34, 2015.
- [73] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [74] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay. Scikit-learn: Machine Learning in Python. . . . *of Machine Learning . . .*, 12:2825–2830, 2012.
- [75] C. A. Pope III and D. W. Dockery. Health effects of fine particulate air pollution: lines that connect. *Journal of the air & waste management association*, 56(6):709–742, 2006.
- [76] PyBossa. Pybossa documentation, 2016. Available at <http://docs.pybossa.com/en/latest/> Accessed: 1 December 2016.
- [77] I. Rimaityte, T. Ruzgas, G. Denafas, V. Racys, and D. Martuzevicius. Application and evaluation of forecasting methods for municipal solid waste generation in an Eastern-European city. *Waste management & research : the journal of the International Solid Wastes and Public Cleansing Association, ISWA*, 30(1):89–98, 2012.
- [78] S. J. Robbins, I. Antonenko, M. R. Kirchoff, C. R. Chapman, C. I. Fassett, R. R. Herrick, K. Singer, M. Zanetti, C. Lehan, D. Huang, et al. The variability of crater identification among expert and community crater analysts. *Icarus*, 234:109–131, 2014.
- [79] R. E. Schapire. The boosting approach to machine learning: An overview. In *Nonlinear estimation and classification*, pages 149–171. Springer, 2003.
- [80] E. Shochat, P. S. Warren, S. H. Faeth, N. E. McIntyre, and D. Hope. From patterns to emerging processes in mechanistic urban ecology. *Trends in ecology & evolution*, 21(4):186–191, 2006.
- [81] A. A. Shusterman, V. E. Teige, A. J. Turner, C. Newman, J. Kim, and R. C. Cohen. The berkeley atmospheric co₂ observation network: initial evaluation. *Atmospheric Chemistry and Physics*, 16(21):13449–13463, 2016.

- [82] E. G. Snyder, T. H. Watkins, P. A. Solomon, E. D. Thoma, R. W. Williams, G. S. Hagler, D. Shelow, D. A. Hindin, V. J. Kilaru, and P. W. Preuss. The changing paradigm of air pollution monitoring, 2013.
- [83] J. A. Tarr. Historical perspectives on Hazardous Wastes in the United States. *Waste Management & Research*, (September 1984):3–41, 1985.
- [84] R. Taylor and A. Allen. Waste disposal and landfill: information needs. 2006.
- [85] P. V. Thakuriah, N. Tilahun, and M. Zellner. *Seeing Cities Through Big Data: Research, Methods and Applications in Urban Informatics*. Springer, 2016.
- [86] I. Tharoor. A huge garbage fire in india’s biggest city was so bad you could see it from space, February 2016. [Online; posted 03-February-2016].
- [87] The US Environmental Protection Agency. *List of Municipal Solid Waste Landfills*. 1996.
- [88] The US Environmental Protection Agency. *Municipal Solid Waste Generation , Recycling , and Disposal in the United States Tables and Figures for 2012*. Number February. 2014.
- [89] United Nations. World Urbanization Prospects: The 2014 Revision, Highlights (ST/ESA/SER.A/352). *New York, United*, page 32, 2014.
- [90] R. Van Haaren and N. Themelis, Nickolas, Goldstein. The state of garbage in America. *Biocycle*, 51(10):16–23, 2010.
- [91] B. H. Wade and R. D. Bullard. Dumping in Dixie: Race, Class and Environmental Quality. *Contemporary Sociology*, 20(6):911, 1991.
- [92] Y. Wang, J. Li, H. Jing, Q. Zhang, J. Jiang, and P. Biswas. Laboratory evaluation and calibration of three low-cost particle sensors for particulate matter measurement. *Aerosol Science and Technology*, 49(11):1063–1077, 2015.
- [93] L. Xu, P. Gao, S. Cui, and C. Liu. A hybrid procedure for MSW generation forecasting at multiple time scales in Xiamen City, China. *Waste Management*, 33(6):1324–1331, 2013.
- [94] M. J. G. Zade and R. Noori. Prediction of municipal solid waste generation by use of artificial neural network: A case study of mashhad. 2007.